Causal effect of changes in industrial air pollution on asthma rates in Western Pennsylvania

by

Margaret Kuzemchak

Bachelor of Science in Chemistry and Earth Science, Pennsylvania State University, 2019

Submitted to the Graduate Faculty of the

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Margaret Kuzemchak

It was defended on

April 25, 2022

and approved by

Eric Roberts, PhD, Department of Health Policy Management

Jeanine Buchanich, PhD, Department of Biostatistics

Jenna Carlson, PhD, Department of Biostatistics

Thesis Advisor: Ada Youk, PhD, Department of Biostatistics

Causal effect of changes in industrial air pollution on asthma rates in Western Pennsylvania

Margaret Kuzemchak, MS

University of Pittsburgh, 2022

This thesis assesses the causal effects of the closure of the Horsehead Holding Corporation zinc smelter in Beaver County in 2014 on asthma encounters through a change in air pollution. The smelter was one of many industrial sources of air pollution in Southwestern Pennsylvania's history. Air pollution, specifically in the form of PM 2.5, has a negative effect on public health by increasing the risk of cardiovascular disease, asthma, and mortality. The empirical methods used in this study were a difference in difference analysis and an instrumental variable analysis. The results show a decrease in PM 2.5 air pollution and total asthma encounters in Beaver County due to the closure. Although the results were not statistically significant, clinical significance may be given to the findings. Both methods concluded that 47 asthma encounters per year were prevented in the area exposed to the zinc smelter due to the closure. Further analysis could lend to the understanding of total community health effects of point source polluters in Southwestern Pennsylvania. This thesis, despite its limitations, provides evidence of a decrease in particulate air pollution and asthma outcomes due to the closure of the smelter that are important for public health in Beaver County and Southwestern Pennsylvania.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

Southwestern Pennsylvania is rooted in a long history of industry. With this industry came many public health concerns such as poor air pollution for the citizens of the region. While many of the sources of this air pollution have moved out of the area, some have remained until the twenty-first century. This includes the Horsehead Holding Corporation's zinc smelter in Beaver County, Pennsylvania. The plant was incorporated by the Horsehead Holding company in 2003 and closed for operation in mid-2014 (Gough, 2016). The previous site of the zinc smelter is currently being developed by Shell Corporation to be a petrochemical "cracker" factory (WTAE, 2013).

Figure 1 Horsehead Holding Corporation zinc smelter in Beaver County, PA (Gough, 2016)

PM 2.5 is particulate matter with a diameter less than 2.5 microns. These particles are inhalable and air pollution in the form of PM 2.5 is known to negatively affect health. The smaller the particulate matter is, the greater risk it poses for public health. Particles less than 10 microns

can penetrate deep into the lungs and even into the blood stream (United States Environmental Protection Agency, 2021). According to the National Ambient Air Quality Standards set by the Environmental Protection Agency, average exposure of 12 $\mu g/m^3$ over a year or 35 $\mu g/m^3$ over 24 hours can be harmful to human health (United States Environmental Protection Agency, 2022). However, the World Health Organization holds higher standards with an annual exposure threshold of 5 $\mu g/m^3$ and a daily threshold of 15 $\mu g/m^3$ (World Health Organization, 2021). Studies have linked PM 2.5 exposure to increased incidence of cardiovascular disease and mortality. There is often a racial disparity in exposures to PM 2.5 due to proximity to industrial sources and higher incidence of the related health effects (Erqou, et al., 2018). PM 2.5 is associated with a decrease in peak expiratory flow in both adults and children with asthma (Edginton, O'Sullivan, King, & Lougheed, 2021).

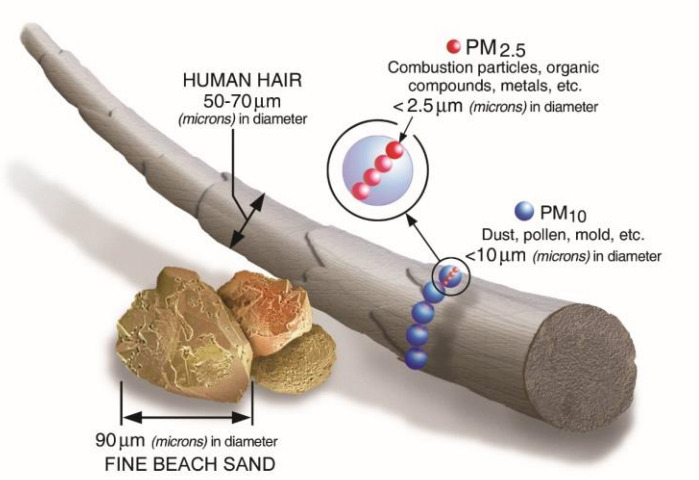

Figure 2 PM Size comparison of PM 2.5 to a human hair

Causal inference methods can be employed when changes in some condition create a natural experiment. This occurs when an event happens that changes the exposure or treatment of one group in a population creating two groups that are practically randomized. Causal studies on

the effect of air pollution on asthma have used sudden changes in a source of air pollution as exogenous source of variation. Recent literature uses difference in difference study designs to model the effect of the shutdowns from the COVID-19 Pandemic on various health outcomes through a decrease in air pollution (Son, et al., 2020).

Very few causal studies have been done to understand the effects of a shutdown of a single point source polluter on health. One study analyzed the effect of a coal power plant closure on asthma outcomes through a change in SO_2 using a difference in difference model. This study showed a decrease in hospitalizations due to asthma in the period after the closure (Casey, et al., 2020). However more research is needed in this area to fully understand how different polluters and pollutants affect public health and how to improve air quality for affected communities. Before this study, there has not been research on the effects of the closure of the Horsehead Zinc smelter in Beaver County. This thesis is an important analysis of the public health impacts of air pollution from industrial sources in Southwestern Pennsylvania.

The closure of the Horsehead zinc smelter created a natural experiment in which one group of the population experiences a sudden change in exposure to PM 2.5. The objective of this thesis is to perform a causal analysis of the closure of the Horsehead zinc smelter in Beaver County on asthma encounters through a change in air pollution. The two empirical methods employed are difference in difference and instrumental variables. We hypothesize that the closure of the zinc smelter will lead to a change in air pollution and asthma encounters for an exposed group of county tracts within Beaver County. The unexposed groups will be used as a counterfactual for the causal analysis; this is a group that shows what trends in air pollution and asthma encounters would have been but for the closure of the smelter.

2.0 Methods

The final data set includes forty-nine county tracts in Beaver County with twenty-four repeated observations covering four quarters over six years (2011 – 2016). Each observation contains an indicator for being in the region exposed to the zinc smelter, an indicator for being in the post-closure period after the smelter closed, an average PM 2.5 value for that county tract in $\mu g/m^3$, and a rate of asthma encounters recorded in that tract during that quarter. The response variable measured at the tract quarter level was the rate of asthma encounters recorded per ten thousand people living in the county tract. Two different response variables were investigated: total asthma encounters and hospitalizations and emergency department visits.

2.1 Data Management

To create the final data set, data from two primary sources were used – University of Pittsburgh Medical Center (UPMC) asthma patient data and PM 2.5 satellite data compiled and hosted by Washington University of St. Louis Atmospheric Composition Analysis Group (Atmospheric Composition Analysis Group, 2022). Population data for Beaver County tracts from the American Community Survey in 2011 – 2016 was utilized as well (United States Census Bureau, 2022).

2.1.1 Air Pollution Data Cleaning

The air pollution data was originally presented as a raster dataset – a pixelized image – of monthly PM 2.5 values for North America from the Washington University of St. Louis Atmospheric Composition Analysis Group. The monthly data for January, April, July, and October were used to represent the four quarters of each year. In R, each raster was clipped to include only Beaver County. Using zonal statistics, a mean PM 2.5 value was calculated for each county tract during each quarter. Zonal statistics is a spatial data analysis tool that calculates an indicated statistic, in this case the mean, from raster data for a specified geographic zone. The air pollution data consisted of raster data with a $0.01° \times 0.01°$ spatial resolution. Each pixel contained a value for PM 2.5. The zonal statistics tool created an average value for each county tract based on the values of the pixels within the tract. These values were compiled into a final spatial dataset that contained the average PM 2.5 value for each of the tracts at every quarter.

2.1.2 Asthma Data Cleaning

The asthma data was pulled from a larger data set including UPMC patient data from multiple counties in Southwestern Pennsylvania using MySQL Workbench. The data pull included patient study ID, patient diagnosis code, latitude and longitude of the patient's home address, an indicator for asthma as the primary diagnosis, and the diagnosis date. The data pull selected patients only in Beaver County by latitude and longitude who had a diagnosis code associated with asthma.

The data was further cleaned in R to select only encounters that took place in 2011 through 2016. Two levels of asthma encounters were analyzed – the total rate of asthma encounters

recorded and the rate of hospitalizations and emergency room (ER) visits. Total asthma encounters were determined by having a primary diagnosis code related to asthma. Hospitalizations and ER visits were a subset of the total asthma encounters and determined by a diagnosis code relating to the emergency department, hospitalizations, ambulance use, or inpatient discharges. Duplicate encounters were defined as a diagnosis that had the same date and patient study ID and were removed from the dataset. Each encounter was labeled by quarter and year. In ArcMap, the encounters were then plotted on the shapefile map of Beaver County that contained the air pollution data for each quarter. Each encounter was assigned a county tract based on the patient's home latitude and longitude. The number of encounters per tract per quarter were counted in R. A rate for each county tract during each quarter was calculated by dividing the total number of asthma encounters in that time and tract by the population of that tract from the American Community Survey for the corresponding year.

2.1.3 Assigning Exposure and Post-Closure Indicators

According to local news sources, the final closure of the zinc smelter was in mid-year 2014 (Gough, 2016). The post-closure indicator was coded for quarters fifteen to twenty-four. An exposed area was determined by looking at the tract averaged air pollution maps in the pre-closure period. The geography of Beaver County was considered when making these regions. Pollution is likely to remain in the river valleys and move from West to East.

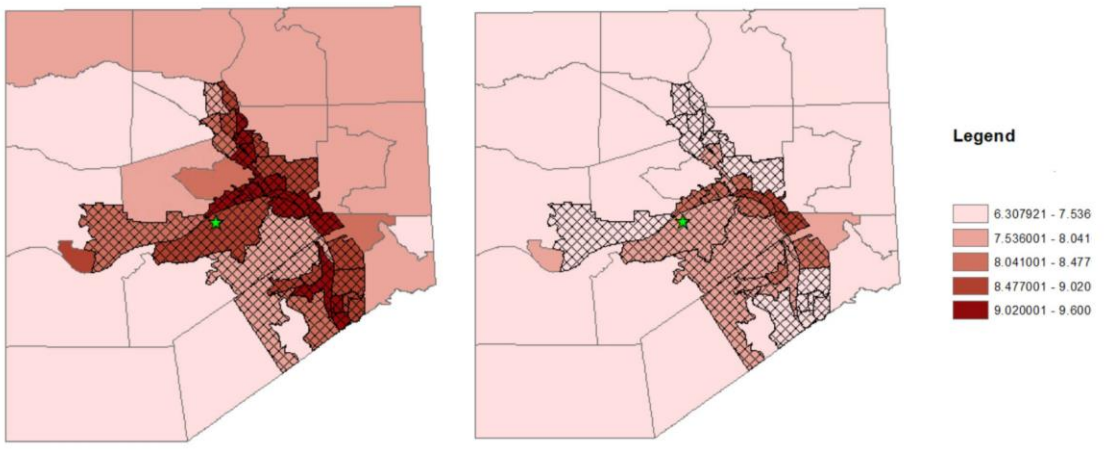

Figure 3 Comparison of Beaver County PM 2.5 values in exposed tracts in October 2011 and October 2015

Each quarterly PM 2.5 map was examined in ArcMap and compared to find an overlap of areas with relatively high PM 2.5 levels near the zinc smelter. The tracts with the highest PM 2.5 values, surrounding the zinc smelter were placed in the exposed set. The exposed tracts are shown cross hatched in the figure above. The figure shows a map from October 2011 and October 2015 of Beaver County with the mean PM 2.5 values of each tract. The star represents the location of the zinc smelter. The maps show an overall decrease in PM 2.5 over the study period.

2.2 Statistical Analysis

The purpose of this thesis is to perform a causal analysis of the effects of a zinc smelter closure in Beaver County on asthma rates. Causal analysis allows for causal assumptions to be made from observational data using a counterfactual. The counterfactual offers a comparison of

what outcome would have occurred but for some intervention of treatment (Angrist & Pischke, 2009). In natural experiments, an instrument is employed to isolate the causal effects of exposures from unmeasured confounders. The instrumental variable does this by using exogenous variation that serves to quasi-randomize groups to the exposure of interest, thereby controlling for unmeasured confounders.

A difference in difference analysis was performed to determine the differential change in asthma rates before and after the closure of the zinc smelter between the exposed and unexposed areas. An instrumental variable analysis was done using two-stage least squares regression (2SLS) to determine the causal effect of the closure on asthma rates through a change in air pollution.

Closure → Air Pollution → Asthma Rates **Equation 1**

The equation above shows the causal pathway for this analysis. The closure was expected to cause a change in asthma rates only through a change in air pollution. The closure acts as the instrument and is assumed to be a random event that is not associated with asthma rates or other unmeasured time-varying confounders. (As discussed below, the difference-in-differences and instrumental variable designs hold constant time-invariant confounders across exposed and unexposed tracts.)

2.2.1 Exploratory Analysis

The final data sets consist of 1176 observations for total asthma encounters and hospitalizations and emergency department visits. The unit of analysis was the tract-quarter. An exploratory analysis of the data was done to show how the closure of the zinc smelter affected PM 2.5 levels.

$$PM2.5_{z\,t} = \beta_0 + \beta_1 * exposed_z + \beta_2 * post_t + \beta_3 * exposed_z * post_t + \beta_4 * summer_t + \varepsilon_{zt} \quad \textbf{Equation 2}$$

The variable $exposed_z$ is the exposure indicator, and $post_t$ is the post closure quarter indicator. The subscript *z* indicates that the data is at the tract level and the subscript *t* indicates that the data is at the quarter of the year level. An indicator for summer is represented by β_4. The response variable is the PM 2.5 value in μg/m^3. The coefficient on the interaction between $exposed_z$ and $post_t$, β_3, is the difference in difference estimator. It is the differential change in PM 2.5 values between the two exposure regions before and after the closure. This is also the first stage model of the 2SLS regression. This relationship was graphed to understand trends in PM 2.5 levels before and after the closure for the two groups. The change in asthma rates for the two exposure zones over time was graphed. This was done for both total asthma encounters and hospitalizations and emergency department visits.

2.2.2 Difference in Difference Analysis

The empirical method for the difference in difference analysis uses the zinc smelter closure as a quasi-randomizer. The difference in difference approach compares the group level trends of an exposure group to a control group before and after a change in exposure. In this method, time invariant confounders are held constant because the model compares the changes over time between the exposure and control groups. Thus, time-invariant differences between the groups will not confound the comparisons of changes across the groups—the focus of the difference in differences design. A key assumption of the difference in difference study design is that the trends in both groups would be the same but for the intervention (Angrist & Pischke, 2009). Practically,

this means the pre-trends must be parallel. The differential change between the two groups before and after the change in exposure is the statistic of interest.

An intention to treat model shows the relationship between the instrument and the outcome. In this study, the intention to treat model estimates the differential change in asthma rates between the exposed and unexposed county tracts before and after the zinc smelter closure. It allows for an estimate of differential change for those who were possibly affected by the exposure, independent of the actual PM 2.5 value.

$$asthma_{z\,t} = \beta_0 + \beta_1 * exposed_z + \beta_2 * post_t + \beta_3 * exposed_{\,z} * post_t + \beta_4 * summer_t + \varepsilon_{zt} \quad \textbf{Equation 3}$$

The coefficient of the interaction between $exposed_z$ and $post_t$, β_3, is the difference in difference estimator. It is the differential change in asthma rates between the two exposure regions before and after the closure.

The difference in difference plots were created with a break at quarter 15 to visualize the different trends for each exposure group before and after the closure of the zinc smelter. This analysis was done for both response variables – total asthma encounter rates and hospitalization and emergency department visit rates.

2.2.3 Instrumental Variable Analysis

The instrumental variable analysis uses the closure of the zinc smelter as an instrument acting on air pollution and asthma encounters. The closure of the zinc smelter quasi-randomizes the exposed and unexposed tracts into unbiased groups – before the closure and after the closure due to exogenous variation. The two main assumptions of an instrumental variable analysis are the exclusion restriction and relevance. The exclusion restriction states that the instrument is only

associated with the outcome through the exposure and not through another pathway with confounding variables (Angrist & Pischke, 2009). The relevance assumption states that the instrument has a significant effect on the exposure. This assumption can be tested by looking at the effect of the instrument in the first stage (Angrist & Pischke, 2009).

In the 2SLS regression model, the first stage is a regression of the exposure on the instrument. The second stage regresses the outcome on the fitted values of the exposure from the first stage. The instrument is not included in the second state because the variation from the instrument is already included in the fitted values of the exposure (Angrist & Pischke, 2009). The local average treatment effect (LATE) is the coefficient on the predicted exposure values in the second stage. Because the exposure is now as good as random from the exogenous variation of the instrument this estimate is unbiased. This number represents the effect of a 1 unit change in PM 2.5 due to the zinc smelter closure on the rate of asthma encounters per 10,000 people per quarter in the exposed tracts.

The first stage of the two-stage model is represented above as the Equation 2. The predicted values for PM 2.5 from this model are then used in the second stage model to relate with asthma rates.

$$asthma_{z\,t} = \beta_0 + \beta_1 * \widehat{PM}_{2.5\,zt} + \beta_2 * post_t + \beta_3 * exposed_z + \beta_4 * summer_t + \varepsilon_{zt} \qquad \textbf{Equation 4}$$

In this equation $\widehat{PM}_{2.5\,zt}$ represents predicted values for PM 2.5 from the first stage model (Equation 2). The second stage model is the regression of asthma rates on the predicted air pollution values and the covariates, $post_t$ and $exposed_z$; it does not include the interaction between $exposed_z$ and $post_t$ which represents the instrumental variable. The LATE is represented by the coefficient on the predicted PM 2.5 values, β_1.

3.0 Results

The sample consisted of 49 county tracts in Beaver County with 24 repeating observations for each tract totaling a sample size of 1176. The outcomes investigated were total rates of asthma encounters per 10,000 people in each quarter where asthma encounters were defined as total encounters and severe encounters meaning a hospital or emergency department visit. The R code for these analyses can be found in Appendix A.

Table 1 Summary Statistics for PM 2.5 values in the Exposure Zones Before the Closure

	Mean PM 2.5 ($\mu g/m^3$)	Standard Deviation
Exposed	10.2	2.44
Unexposed	8.89	2.38

Before the closure of the zinc smelter, PM 2.5 was higher on average in the exposed tracts than the unexposed tracts by over 1 $\mu g/m^3$. This is expected as the smelter would create higher levels of PM 2.5 in the surrounding area.

Table 2 Summary Statistics for Total Asthma Encounters in the Exposure Zones Before the Closure

	Mean asthma rate per 10,000 people	Standard Deviation
Exposed	4.77	6.86
Unexposed	6.67	8.13

The rate of total asthma encounters in the pre-closure period is higher on average in the unexposed tracts compared to the exposed tracts. This does not mean that there is no relationship between PM 2.5 and asthma encounters. Individuals in the unexposed tracts may have time-invariant demographics that make those populations have higher rates of asthma such as age, race, and income level. This will not be an issue for the models used in this study as they will hold these time-invariant confounders constant and compare the differential changes in the exposure groups before and after the closure. The study design is validated by this relationship as it would be difficult to draw conclusions about how air pollution and the smelter affect asthma rates in Beaver County without an exogenous source of variation.

Table 3 Summary Statistics for Hospitalizations and ER visits in the Exposure Zones Before the Closure

	Mean asthma rate per 10,000 people	Standard Deviation
Exposed	0.60	1.63
Unexposed	0.54	1.36

The rate of asthma hospitalizations and ER visits in the exposed and unexposed zones before the closure are very low in both the exposure groups. The large standard deviations in the rates of both asthma outcomes show that there is a large spread of data about the mean.

3.1 Exploratory Analysis

Exploratory analysis was done to assess temporal trends of both PM 2.5 and asthma exacerbations in both exposure groups.

3.1.1 First Stage Relationship

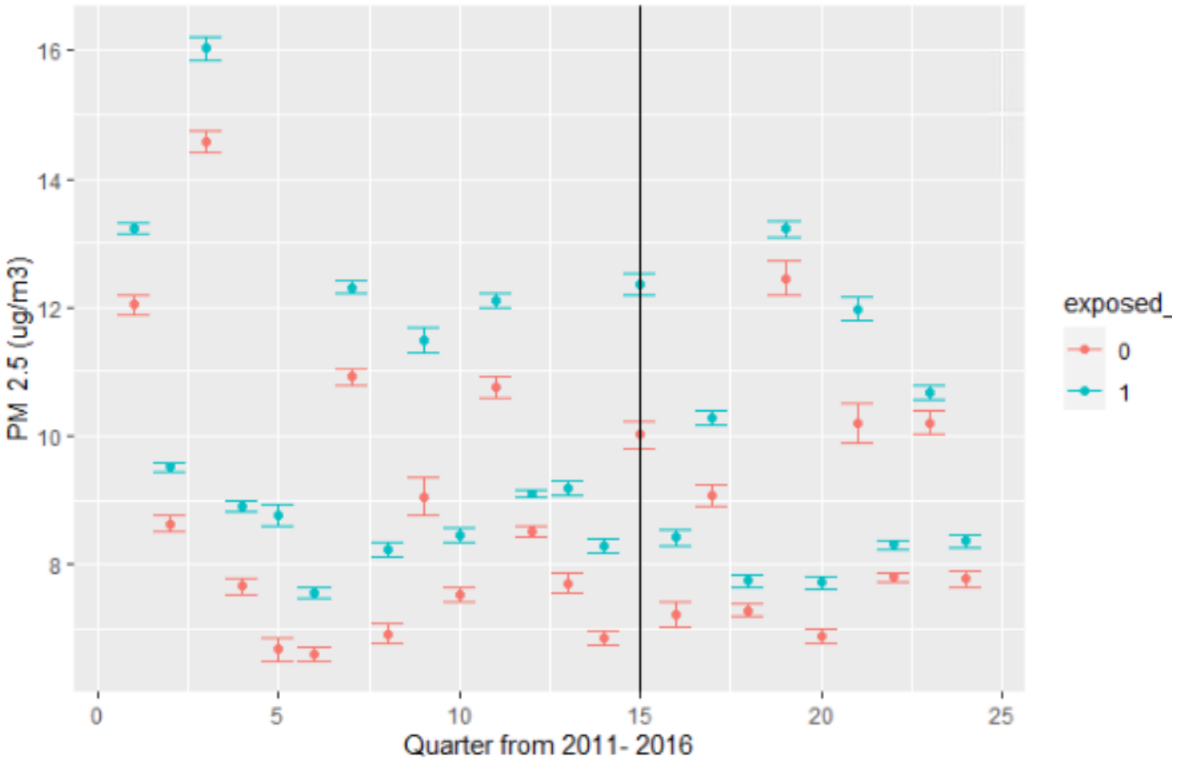

Figure 4 Average PM 2.5 for exposed and unexposed tracts from 2011 - 2016

The graph above displays the changes in average PM 2.5 for county tracts in each exposure group from 2011 – 2016. An exposure value of 0 corresponds to unexposed county tracts and an exposure value of 1 corresponds to exposed county tracts. The vertical line represents the threshold between the pre-closure and post-closure period. Specifically, it represents that the smelter closed mid-year in 2014.

Figure 5 Temporal Trends of PM 2.5 for exposued and enexposed groups

The graph above displays the temporal trend in PM 2.5 for county tracts in each exposure groups over time from 2011 – 2016. An exposure value of 0 corresponds to unexposed county tracts and an exposure value of 1 corresponds to exposed county tracts. Figures 4 and 5 visually represent the first stage of the 2SLS model. There is an overall decrease in PM 2.5 over the study period. In the post-closure period, PM 2.5 increased slightly in the exposed region and more sharply in the unexposed region.

3.1.2 Intention to Treat

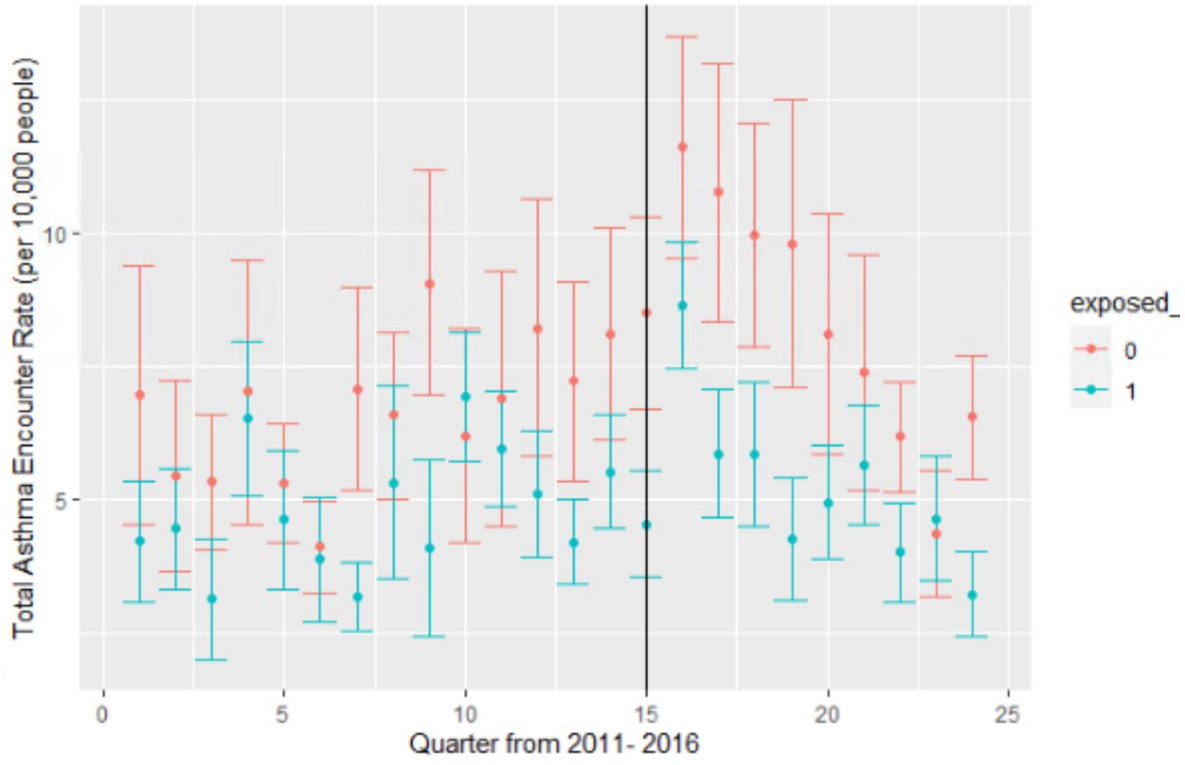

Figure 6 Total Astham Encounters from 2011 - 2016 in exposed and unexposed tracts

The graph above visualizes the relationship between rates of total asthma encounters per 10,000 people and time by quarters for the two exposure groups. Rates of total asthma encounters appear to be higher in the unexposed group compared to the exposed group.

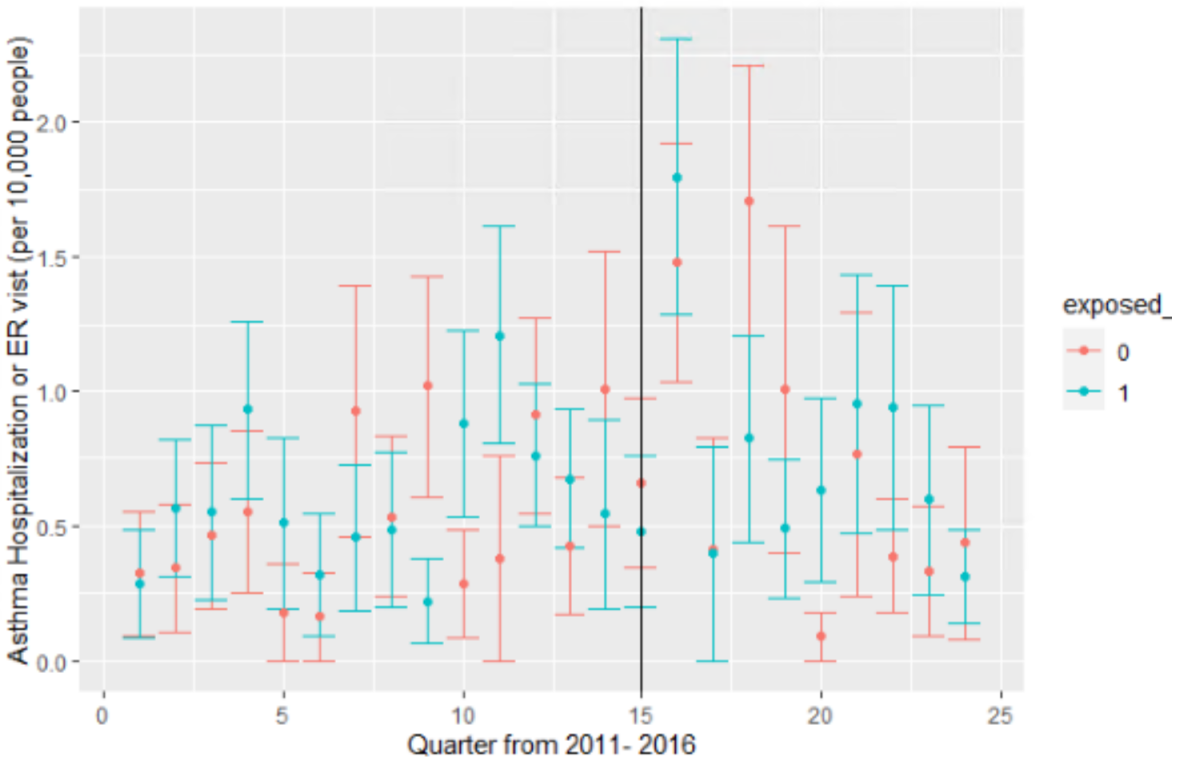

Figure 7 Event Study of Hospitalizations and ER visits from 2011 - 2016 in exposed and unexposed tracts

The graph above visualizes the relationship between rates of hospital and emergency room visits per 10,000 people and time by quarters for the two exposure groups. The rates of hospitalizations appear to be similar in the two exposure groups. The difference in difference analysis assesses whether the differential change between the two groups is significant for each outcome.

3.2 Difference in Difference Analysis

A difference in difference analysis was performed to assess the differential change in asthma rates before and after the zinc smelter closure.

3.2.1 Intention to treat model

An intention to treat model was performed to assess the effect of the zinc smelter closure on asthma encounters. This model is represented by Equation 3.

Table 4 Estimates for Intention to treat model for Total Asthma Encounters per 10,000 people

Coefficient	Estimate	P-value	Std. Error	95% CI
Intercept, β_0	6.85	< 2e-16 ***	0.468	(5.93, 7.77)
Exposed, β_1	-1.89	0.001	0.574	(-3.02, 0.77)
Post, β_2	1.72	0.015 ***	0.708	(0.33, 3.11)
Instrument, β_3	-1.28	0.151	0.889	(-3.02, 0.46)
Summer, β_4	-0.843	0.086	0.490	(-1.80, 0.12)

*** indicates statistical significance at the 0.05 level

The difference in difference estimator is represented by β_3. For rates of total asthma encounters, the difference in difference estimator is -1.28 and has a p-value of 0.151. The decrease in the rate of asthma encounters after the closure for the exposed tracts was 1.28 greater per 10,000 people per quarter than in the unexposed tracts.

Figure 8 Difference in Differnce for Total Asthma Encounters

The graph above displays the trends in average rate of total asthma encounters in the exposed and unexposed tracts over time. The rates of total asthma encounters are increasing in the pre-closure period in a similar trajectory for both exposure groups. Parallel trends in the pre-closure period are assessed in the discussion and tested in Appendix B. After the closure, the total rates of asthma encounters are decreasing in both exposure groups. A negative difference in difference estimator from Table 4 shows that there is a greater decrease in total asthma encounter rates per 10,000 people in the exposed tracts before vs. after the closure compared to contemporaneous changes in the unexposed tracts.

Table 5 Estimates for Intention to Treat Model for Hospitalizations and ER Visits per 10,000 people

Coefficient	Estimate	P-value	Std. Error	95% CI
Intercept, β_0	0.54	3.93e-5 ***	0.122	(0.30, 0.78)
Exposed, β_1	0.063	0.647	0.137	(-0.21, 0.33)
Post, β_2	0.19	0.253	0.169	(-0.14, 0.52)
Instrument, β_3	-0.037	0.818	0.212	(-0.45, 0.38)
Summer, β_4	-0.049	0.755	0.117	(-0.28, 0.18)

*** indicates statistical significance at the 0.05 level

For rates of hospitalizations and ER visits, the difference in difference estimator is -0.037 and has a p-value of 0.818. The decrease in the rate of asthma hospitalizations after the closure for the exposed tracts was 0.037 greater per 10,000 people per quarter than in the unexposed tracts.

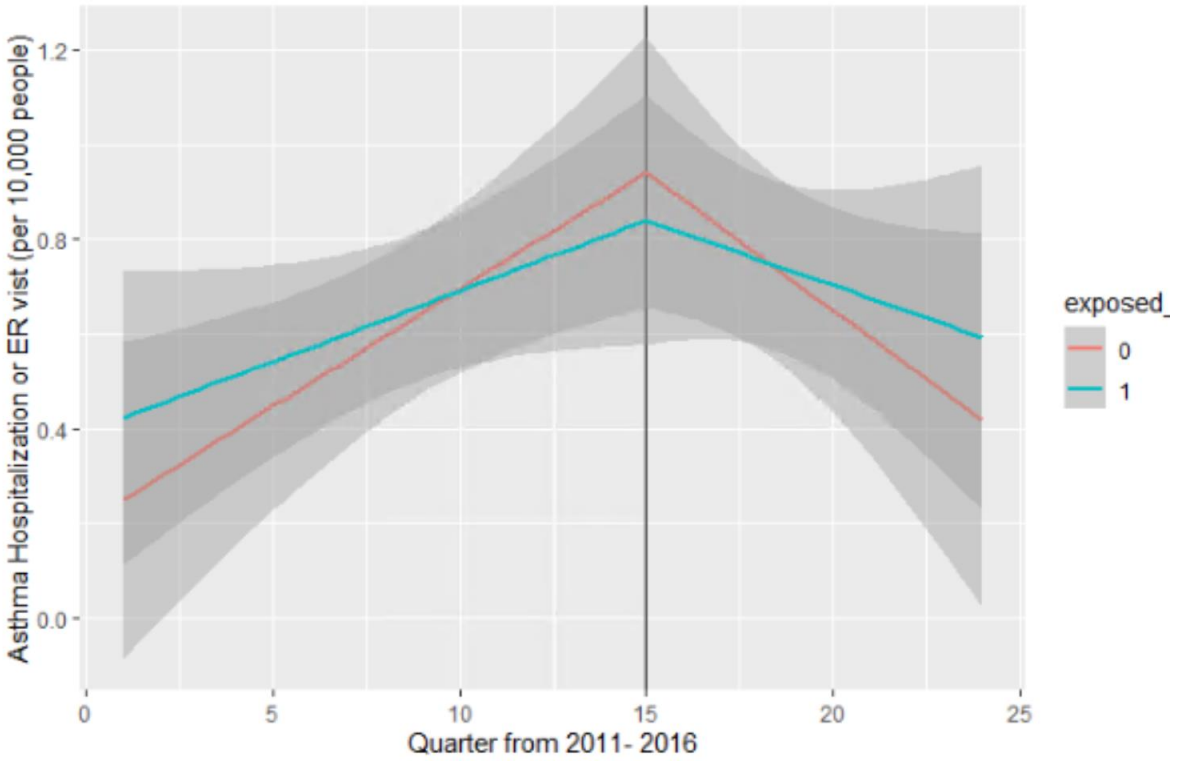

Figure 9 Difference in Differnce for Hospitalizations and Emergency Room visits

The graph above displays the trends in average rate of hospitalizations and emergency room visits in the exposed and unexposed tracts over time. The rates of hospitalizations and emergency room visits increase in the pre-closure period and decrease in the post-closure period. Although the direction of the relationship is the same, the difference in difference estimator is much smaller for hospitalizations and emergency room visits compared to total asthma encounters.

3.3 Instrumental Variable Analysis

An instrumental variable analysis was performed to assess the causal effects of the zinc smelter closure on asthma rates through a change in air pollution.

3.3.1 First Stage Model

The first stage model represents the first part of the causal pathway – the effect of the zinc smelter closure on PM 2.5 air pollution. This model is represented by Equation 2.

Table 6 First Stage Model Estimates Effects of Closure on Air Pollution

Coefficient	Estimate	P-value	Std. Error	95% CI
Intercept, β_0	8.12	<2e-16 ***	0.106	(7.91, 8.33)
Exposed, β_1	1.34	<2e-16 ***	0.130	(1.09, 1.59)
Post, β_2	-0.31	0.055	0.160	(-0.62, 0.004)
Interaction, β_3	-0.32	0.109	0.201	(-0.71, 0.07)
Summer, β_4	3.60	<2e-16 ***	0.111	(3.38, 3.82)

*** indicates statistical significance at the 0.05 level

The differential change in PM 2.5 between the two exposure groups is -0.32 $\mu g/m^3$. The decrease in PM 2.5 after the closure in the exposed tracts was 0.32 $\mu g/m^3$ greater per quarter than in the unexposed tracts. This is consistent with trends in Figure 5. The p-value of 0.109 shows that the closure is a moderately strong instrument for air pollution.

3.3.2 Second Stage Model

The second stage model represents the second part of the causal pathway – the effect of air pollution on asthma rates. This model is represented by Equation 4. The predicted values of PM 2.5 from the first stage model were used in the second stage model.

Table 7 Second Stage Estimates for Total Asthma Encounters per 10,000 people

Coefficient	Estimate	P-value	Std. Error	95% CI
Intercept, β_0	-25.38	0.263	22.64	(-69.8, 19.0)
$\widehat{PM}2.5$, β_1	3.97	0.151	2.76	(-1.44, 9.38)
Post, β_2	2.94	0.046 ***	1.48	(0.039, 5.84)
Exposure, β_3	-7.20	0.031 ***	3.35	(-13.8, -0.63)
Summer, β_4	-15.12	0.128	9.93	(-34.6, 4.34)

*** indicates statistical significance at the 0.05 level

The β_1 value represents the local average treatment effect. The LATE for total asthma encounters is 3.97 with a p-value of 0.151. For a one unit decrease in PM 2.5, there is a decrease of 3.97 total encounters per 10,000 people per quarter in the exposed tracts due to the closure of the zinc smelter.

Table 8 Second Stage Estimates for Hospitalizations and Emergency Room Visits per 10,000 people

Coefficient	Estimate	P-value	Std. Error	95% CI
Intercept, β_0	-0.68	0.889	5.414	(-11.3, 9.93)
$\widehat{PM}2.5$, β_1	0.15	0.818	0.660	(-1.14, 1.44)
Post, β_2	0.24	0.495	0.325	(-0.40, 0.88)
Exposure, β_3	-0.14	0.861	0.800	(-1.71, 1.43)
Summer, β_4	-0.58	0.806	2.38	(-5.24, 4.08)

*** indicates statistical significance at the 0.05 level

The LATE for total hospitalizations and emergency room visits is 0.15 with a p-value of 0.818. For a one unit decrease in PM 2.5, there is a decrease of 1.5 hospitalizations and ER visits

per 100,000 people per quarter in the exposed tracts due to the closure of the zinc smelter (0.15 hospitalizations and ER visits per 10,000 people per quarter = 1.5 per 100,000 people per quarter).

4.0 Discussion

4.1 Interpretation of Results

The results from the difference in difference analysis and the two-stage least squared regression model can be applied to meaningful measured differences in the context of the study. In general, across both methods, the rate of total asthma encounters had stronger relationships with air pollution changes and the closure of the zinc smelter compared to hospitalizations and ER visits. This may be due to much lower rates of hospitalizations and ER visits across all county tracts.

In the difference in difference analysis, the results show that the average rates of total asthma encounters decreased by 1.28 more per 10,000 people per quarter in the exposed tracts compared to the unexposed tracts after the closure of the zinc smelter and the average rate of hospitalizations decreased by 0.037 per 10,000 people per quarter. While neither of these results are statistically significant, there may be clinical significance to the decrease in the rate of total asthma encounters. There are around 92,000 people who live in the exposed region of Beaver County in any given year (United States Census Bureau, 2022). The difference in difference estimator shows that after the closure there were 47 fewer total asthma encounters per year in the exposed area compared to the unexposed area. This was calculated by dividing the rate per 10,000 people by 10,000 and multiplying by 92,000, the total population in the exposed tracts, then multiplying by 4, the number of quarters per year.

Both the first and second stage in the two-stage least squares analysis for the instrumental variable allow for relevant results and interpretation. The first stage model shows that there is a

greater decrease in PM 2.5 values in the exposed tracts after the closure compared to the unexposed tracts. This change can be seen visually in Figure 3 and Figure 5. In Figure 3, comparisons of the same month in 2011 and 2015 shows that overall air pollution has decreased, and the exposure zone holds less tracts with higher relative levels of air pollution. The decrease in PM 2.5 per quarter for the exposed tracts after the closure was 0.32 $\mu g/m^3$ greater than in the unexposed tracts. The direction of the relationship – that air pollution decreased more in the exposed regions after the closure – was as expected. However, the magnitude of the relationship was less than expected. The p-value on the difference in difference estimator, the instrument, in the first stage model was .109, which may be considered weak. This means that the model is about 90% confident that there was an effect larger than 0 of the zinc smelter closure on the air pollution. Simply, the zinc smelter closure may not be strongly related with the PM 2.5 values between the chosen exposure regions. This challenges the relevance assumption of instrumental variables that the instrument strongly predicts the exposure. This is rooted in the limitation that the exposure zone cannot be accurately portrayed as a set of county tracts. More information on this limitation is explained below.

In the second stage least squares model, the predicted PM 2.5 values from the first stage regression are used to model the rate of asthma encounters due to the decrease in air pollution because of the closure of the zinc smelter. The results show that a one unit decrease in PM2.5, leads to a decrease of 3.97 per 10,000 people in total asthma encounters in the exposed tracts per quarter due to the closure of the zinc smelter and a decrease of 1.5 per 100,000 people for hospitalizations and ER visits. Given the population of the exposed tracts, the LATE values equate to 146 asthma encounters prevented per year and 3 hospitalizations and ER visits prevent per two years due to a decrease in PM 2.5 of 1 $\mu g/m^3$ per quarter. This was calculated by dividing the LATE by 10,000 then multiplying by 92,000, the total population in the exposed tracts, and

multiplying by 4, the number of quarters per year. The average decrease of 0.32 $\mu g/m^3$ per quarter due to the closure of the zinc smelter from the first stage model prevented about 47 asthma encounters per year. This was calculated by multiplying 146, calculated above by the difference in difference estimator from the first stage, 0.32. Again, this number is not statistically significant, but may represent clinical significance, especially for a rural county.

4.2 Confounding and Limitations

The largest threat to internal validity in the empirical methods used in this study is unmeasured time varying confounding. An example of time varying confounding would occur if changes in the distribution of demographic measures differed within the exposed or unexposed before and after the closure. Because the study size is only one rural county, it is plausible that there are common shocks among the entire county after the closure of the zinc smelter. This supports the assumption that there is no unmeasured time varying confounding. However, a potential violation of this assumption could occur if there were changes in other polluting sources—unrelated to the smelter closure—that differentially affected the exposed and control tracts in this analysis. Furthermore, in future analyses, demographic indicators such as race, age, and poverty level, if available, could have been used to increase precision of the estimates and narrow the confidence intervals.

Tests were performed to assess the validity of the assumption on the difference in difference analysis that the pre-closure trends are parallel. The results, found in Appendix B, show that when the model was run not holding time invariant confounders constant, there was no difference in slopes for the exposed and unexposed groups. This indicates that the pre-trends for

asthma encounters and hospitalizations and ER visits are parallel and that there are not unmeasured time varying confounders between exposure groups. However, due to issues of power throughout the entire study, this test may not actually translate to no levels of unmeasured confounding between the exposure groups. An interrupted time series comparison may serve to further understand this relationship an assess confounding.

The results from the 2SLS regression could be conflated due to the moderate strength of the instrument. If the instrument is not strong, there is a low rate of compliance in the exposed groups. (By compliance, we mean that the change in air pollution is estimated to be due solely to the zinc smelter closure and not to other factors.) In this study, an example of this is where county tracts in the exposed groups do not have a high PM 2.5 value in the pre-closure period. The smaller the rate of compliance, the larger the LATE because the entire effect on the outcome is assigned to a small proportion of the sample. This is a problematic relationship and explains why the relevance assumption is important for interpreting the LATE.

Assigning an exposure region for the zinc smelter was complicated and difficult. Data are not readily available for days or periods of high production at the zinc smelter or weather patterns that would have affected the way the pollution settles in the area. Air pollution is subject to wind, temperature, humidity, and other weather patterns. An indicator for the summer was added to the model to attempt to adjust for this variability because PM 2.5 is typically highest in the summer. However, this does not account for monthly or even daily variation of air pollution due to complex weather systems. Some days, the pollution from the zinc smelter may have lingered only in a small area surrounding the site and others it may have been blown far away to neighboring county tracts. It is possible that during some days or months there was low production at the zinc smelter and the pollution levels were low. Because of this, it is impossible to create a perfect exposure zone for

the zinc smelter. This likely contributed to the insignificant difference in difference estimator for the intention to treat model and the first stage model.

Assigning the exposure was also difficult because the data was at the county tract level. A tract was either exposed or not exposed to the smelter. It is likely that some county tracts had only portions of area that were exposed to the zinc smelter. The small number of county tracts used (49) made the analysis very sensitive to the exposure tracts assigned. The addition and removal of certain tracts on the border of the exposure zone yielded very different results.

A solution to these limitations would have been to use person-level data instead of county tract level. More specific zones would not need to follow arbitrary county tract lines and could allow for a more accurate exposure zone. With person level data more observations would occur in each quarter increasing the statistical power of the model. A washout zone could have been used in areas where pollution is inconsistent, and the exposed group could have been a much smaller area where pollution from the zinc smelter is more constant. A future study could reveal if this would yield more significant results.

Another solution attempted was to create a time varying exposure zone. From the "at-risk" set, the quarterly tracts in the pre-closure period with a PM 2.5 value greater than 12 $\mu g/m^3$ were placed in the exposure set. Because of this, the exposure set was time varying and changed temporally throughout the dataset prior to the closure of the zinc smelter. This created a very strong instrument in the first stage model. Some quarters did not have any tracts with a PM 2.5 value above 12 $\mu g/m^3$ and therefore did not have an exposed region at that time. However, a time varying exposure zone possibly introduces other sources of time varying confounding. This method increased instrument relevance but possibly violates the exclusion restriction. By changing

the groups that are in the exposure zone, the assumption that the characteristics of the exposed group and the unexposed groups do not change over time is more difficult to make.

The two methods, with time varying and time invariant exposures, that balance between the two assumptions of instrumental variables, found like results. The difference in difference estimates and the 2SLS regression with the time varying exposure found that 55 total asthma encounters were prevented per year due to the closure of the smelter. These results are very similar to the results presented above for the time invariant exposure zone with 47 total asthma encounters prevented per year due to the closure. This increases our confidence in the results we found above representing the true relationship between the closure and asthma rates. The difference in difference estimator for the first stage was much larger at -3.88 and this is expected as this method of exposure assignment was used to create a strong instrument. The supplementary results can be found in Appendix C.

It is unknown if air pollution and asthma encounters have a linear relationship. It is possible that any exposure above a certain PM 2.5 value can contribute to health complications. Almost all the county tracts in this study from 2011-2016 exceed the annual threshold of 5 $\mu g/m^3$ set by the World Health Organization (World Health Organization, 2021). This could mean that the effect of the closure was diluted because there is already too much ambient pollution in the area.

The asthma patient data was only from UPMC and was assumed to represent the rate of asthma encounters for each tract for the entire population of Beaver County. It is unknown if other providers in the area had the same proportions of asthma encounters. Patients may not have spent most of their time at home and could have had other levels of exposure from work or school. Furthermore, the address of the patients are the latitude and longitude of their home address as of 2021 not necessarily their historical addresses in the study period of 2011 – 2016. The diagnosis

codes were difficult to interpret and without a data dictionary it is unknown if all the codes for hospitalizations and ER visits were correctly identified. This could have contributed to weak associations between asthma hospitalizations and ER visits with the closure and PM 2.5 values. The satellite air pollution data was from January, April, July and October for 2011 – 2016. The monthly data was then used to represent the entire quarter of the year following that month. A follow up analysis could look at monthly changes as opposed to quarterly.

5.0 Conclusion

This study shows that there was a decrease in air pollution and total asthma encounters in the exposed areas following the closure of the Horsehead Holding Corporation Zinc Smelter in Beaver County in 2014. In Southwestern Pennsylvania from 2011 – 2016 air pollution levels were improving due to increases in environmental protection policies. Air pollution events with high PM 2.5 likely contribute to increased asthma encounters. According to Figure 4, air pollution in the exposed region before the closure of the smelter had multiple monthly averages above 12 $\mu g/m^3$, with the highest average around 16 $\mu g/m^3$. Although there is no monthly threshold for PM 2.5 averages, this likely exceeds the World Health Organization's cutoff for healthy PM 2.5 levels. After the closure of the zinc smelter only two months exceeded 12 $\mu g/m^3$ and most of the other months hovered around 8 $\mu g/m^3$. From Figure 5, the levels of air pollution converge in the exposed and unexposed tracts at the end of the study period. Simply put, the exposure zones have similar levels of air pollution compared to the unexposed zones 2 years after the closure. The difference in difference estimator in the first stage model showed a decrease of 0.32 $\mu g/m^3$ per quarter or 1.28 $\mu g/m^3$ per year of PM 2.5 levels from the zinc smelter. This decrease plays a major role in improving air quality for Beaver County and contributed to the decrease in total asthma encounters in exposed areas.

Air pollution affects a variety of health outcomes including birth defects, cardiovascular disease, and mortality. This analysis could be done for a variety of health effects to see the full picture of how the closure of the zinc smelter affected Beaver County. This analysis could be done for different demographic groups as well such as racial and ethnic groups, age groups, or income level. This would allow for discussion on environmental justice issues with point source polluters.

This research adds to the limited literature on the use of the closure of a point source polluter as an instrument. This method could be applicable to many more studies in Southwestern Pennsylvania. Other closures since the end of the study period include Shenango Coke Works and Cheswick Generating Station, the last coal fired power plant in Allegheny County. This method can be applied to the introduction of an industrial source as well such as the Shell "cracker" plant, located at the same site as the Horsehead zinc smelter in Beaver County. Hydraulic fracturing of the Marcellus Shale has emerged over the entire state in the last 20 years. Difference in difference analysis could be useful in assessing health changes in communities suddenly exposed to fracking.

This study found that the Beaver County tracts that were exposed to the Horsehead Holding Corporation's zinc smelter experienced a greater decrease in air pollution and total asthma encounters after the closure compared to the unexposed tracts. Both the difference in difference analysis and the instrumental variable analysis found that 47 asthma encounters per year were prevented in the exposed area due to the closure of the smelter in 2014. While this result is not statistically significant, we argue it is clinically significant for community health in Beaver County. Although there were limitations to this study that affect the interpretability of the results, there is still evidence of real changes in air pollution and asthma outcomes that are meaningful to public health in Beaver County. The closure of the zinc smelter was a step towards cleaner and healthier air in Southwestern Pennsylvania and this analysis is a piece of the puzzle for understanding the effects of point source polluters on air pollution and public health.

Appendix A Code for Data Cleaning and Statistical Analysis

Appendix A.1 Asthma Patient Data Cleaning

The following code was used to clean the asthma patient data.

#load library

```{r}
library(tidyverse)
##install.packages("plyr")
library(plyr)
```

load data

```{r}
data<- read.csv("asthma_encounters_beaver.csv")
```

This data is the selected data from the SQL editor and includes: just Beaver county patients (by lat long), just asthma DX_CODE, all asthma encounters within those two categories.

HOSPITALIZATIONS AND ER VISITS

unique encounter values

```{r}
unique(data$ENC_TYPE)
```

Above are all of the unique values for encounter types in the data set. A severe exacerbation is defined as a hospital or ER visit. Below, the data is selected for only encounters that are in the hospital, er, or had an ambulance involved. However, some encounter codes above are vague, coded as a number, or left blank. I need to discuss with Jen which I should include as to not bias the dataset.

select encounter types from hospital or ER

```{r}
hosp_er1 <- data %>% filter((ENC_TYPE) == "HOSPITAL-ENCOUNTER" | (ENC_TYPE) == "HOSPITAL ENCOUNTER" | (ENC_TYPE) =="HOSPITAL RESERVATION" | (ENC_TYPE) == "EXTERNAL HOSPITAL ADMISSION" | (ENC_TYPE) == "UPMC HOSPITAL SUMMARY" | (ENC_TYPE) == "EMERGENCY" | (ENC_TYPE) == "ER REPORT" | (ENC_TYPE) == "INPATIENT" )
```

select only encounters in date range from 2011 to 2016

```{r}
hosp_er1 <- hosp_er1 %>%
filter(str_detect(DX_FROM_DATE, '2011|2012|2013|2014|2015|2016'))
```

The study period if from 2011 to 2016 so only encounters within this time period are included

select only primary diagnosis encounters in date range from 2011 to 2016

```{r}
hosp_er1 <- hosp_er1 %>%
filter(str_detect(PRIMARY_DX_IND, 'Y'))
```

remove encounters that had duplicate day and patient id

```{r}
# remove full duplicates
hosp_er1 <- distinct(hosp_er1)
# create subset of data with just patient id and date
hosp_er_no_dups <- select(hosp_er1, STUDY_ID, DX_FROM_DATE, Concat_Latitude_Complete, Concat_Longitude_Complete)
#remove duplicates form that data set
hosp_er <- distinct(hosp_er_no_dups)
```

Some of the data had duplicate entries or had two different entries for one encounter. For example a hospital entry and an ER entry. Observations were removed that had the same date and patient id for separate entries.

The final data set contains the patient ID, date, and location (lat, lon) for each exacerbation

sort encounters into quarters

The study design has encounters grouped by quarter for the year (ex. Jan-Mar 2011)

```{r}
#2011
# 2011 q1
```

```
q1_11 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2011-01|2011-02|2011-03'))
# 2011 q2
q2_11 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2011-04|2011-05|2011-06'))
# 2011 q3
q3_11 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2011-07|2011-08|2011-09'))
# 2011 q4
q4_11 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2011-10|2011-11|2011-12'))
#2012
# 2012 q1
q1_12 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2012-01|2012-02|2012-03'))
# 2012 q2
q2_12 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2012-04|2012-05|2012-06'))
# 2012 q3
q3_12 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2012-07|2012-08|2012-09'))
# 2012 q4
q4_12 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2012-10|2012-11|2012-12'))
#2013
# 2013 q1
q1_13 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2013-01|2013-02|2013-03'))
# 2013 q2
q2_13 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2013-04|2013-05|2013-06'))
# 2013 q3
q3_13 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2013-07|2013-08|2013-09'))
```

```
# 2013 q4
q4_13 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2013-10|2013-11|2013-12'))
#2014
# 2014 q1
q1_14 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2014-01|2014-02|2014-03'))
# 2014 q2
q2_14 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2014-04|2014-05|2014-06'))
# 2014 q3
q3_14 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2014-07|2014-08|2014-09'))
# 2014 q4
q4_14 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2014-10|2014-11|2014-12'))
#2015
# 2015 q1
q1_15 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2015-01|2015-02|2015-03'))
# 2015 q2
q2_15 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2015-04|2015-05|2015-06'))
# 2015 q3
q3_15 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2015-07|2015-08|2015-09'))
# 2015 q4
q4_15 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2015-10|2015-11|2015-12'))
#2016
# 2016 q1
q1_16 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2016-01|2016-02|2016-03'))
```

```
# 2016 q2
q2_16 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2016-04|2016-05|2016-06'))
# 2016 q3
q3_16 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2016-07|2016-08|2016-09'))
# 2016 q4
q4_16 <- hosp_er %>% filter(str_detect(DX_FROM_DATE, '2016-10|2016-11|2016-12'))
```

count number of excerbations in each quarter

```{r}
quarter = c("11-1", "11-2", "11-3", "11-4", "12-1", "12-2", "12-3", "12-4", "13-1", "13-2", "13-3", "13-4", "14-1", "14-2", "14-3", "14-4", "15-1", "15-2", "15-3", "15-4", "16-1", "16-2", "16-3", "16-4")
 run_qrt = c (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 , 16, 17, 18, 19, 20, 21, 22, 23, 24)
counts = c(nrow(q1_11), nrow(q2_11), nrow(q3_11), nrow(q4_11), nrow(q1_12), nrow(q2_12), nrow(q3_12), nrow(q4_12), nrow(q1_13), nrow(q2_13), nrow(q3_13), nrow(q4_13),nrow(q1_14), nrow(q2_14), nrow(q3_14), nrow(q4_14), nrow(q1_15), nrow(q2_15), nrow(q3_15), nrow(q4_15),nrow(q1_16), nrow(q2_16), nrow(q3_16), nrow(q4_16))
asthma_hosper_count <- data.frame(quarter, run_qrt, counts)
```

plot total asthma exacerbations over time

```{r}
plot(asthma_hosper_count$run_qrt, asthma_hosper_count$counts)
```

```

This plot shows how asthma exacerbations change over time. The observed result is unexpected as the Zinc smelter closed in 2013 and cases were expected to decrease after that. The difference in difference analysis will show asthma decreased or increased less in the exposed areas with time.

# export csv

```{r}
write.csv(asthma_hosper_count,"asthma_hosper_count.csv", row.names = TRUE)
write.csv(hosp_er,"asthma_hosper_data.csv", row.names = TRUE)
```

ALL ASTHMA ENCONTERS

An asthma exacerbation could also be defined as any asthma encounter. This would remove bias from the possibility that encounter type codes changed throughout the study time period or reporting got more specific.

select only encounters in date range from 2011 to 2016

```{r}
all_enc1 <- data %>%
filter(str_detect(DX_FROM_DATE, '2011|2012|2013|2014|2015|2016'))
```

The study period is from 2011 to 2016 so only encounters within this time period are included

select only primary diagnosis encounters in date range from 2011 to 2016

```{r}
all_enc1 <- all_enc1 %>%

filter(str_detect(PRIMARY_DX_IND, 'Y'))

```

remove encounters that had duplicate day and patient id

```{r}

# remove full duplicates

all_enc1 <- distinct(all_enc1)

# create subset of data with just patient id and date

all_enc_no_dups <- select(all_enc1, STUDY_ID, DX_FROM_DATE, Concat_Latitude_Complete, Concat_Longitude_Complete)

#remove duplicates form that data set

all_enc <- distinct(all_enc_no_dups)

```

Some of the data had duplicate entries or had two different entries for one encounter. For example a hospital entry and an ER entry. Observations were removed that had the same date and patient id for separate entries.

The final data set contains the patient ID, date, and location (lat, lon) for each exacerbation

sort encounters into quarters

The study design has encounters grouped by quarter for the year (ex. Jan-Mar 2011)

```{r}

#2011

# 2011 q1

q1_11 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2011-01|2011-02|2011-03'))

# 2011 q2

```
q2_11 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2011-04|2011-05|2011-06'))
# 2011 q3
q3_11 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2011-07|2011-08|2011-09'))
# 2011 q4
q4_11 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2011-10|2011-11|2011-12'))
#2012
# 2012 q1
q1_12 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2012-01|2012-02|2012-03'))
# 2012 q2
q2_12 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2012-04|2012-05|2012-06'))
# 2012 q3
q3_12 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2012-07|2012-08|2012-09'))
# 2012 q4
q4_12 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2012-10|2012-11|2012-12'))
#2013
# 2013 q1
q1_13 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2013-01|2013-02|2013-03'))
# 2013 q2
q2_13 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2013-04|2013-05|2013-06'))
# 2013 q3
q3_13 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2013-07|2013-08|2013-09'))
# 2013 q4
q4_13 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2013-10|2013-11|2013-12'))
```

```
#2014

# 2014 q1

q1_14 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2014-01|2014-02|2014-03'))

# 2014 q2

q2_14 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2014-04|2014-05|2014-06'))

# 2014 q3

q3_14 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2014-07|2014-08|2014-09'))

# 2014 q4

q4_14 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2014-10|2014-11|2014-12'))

#2015

# 2015 q1

q1_15 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2015-01|2015-02|2015-03'))

# 2015 q2

q2_15 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2015-04|2015-05|2015-06'))

# 2015 q3

q3_15 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2015-07|2015-08|2015-09'))

# 2015 q4

q4_15 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2015-10|2015-11|2015-12'))

#2016

# 2016 q1

q1_16 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2016-01|2016-02|2016-03'))

# 2016 q2

q2_16 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2016-04|2016-05|2016-06'))
```

```
# 2016 q3
q3_16 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2016-07|2016-08|2016-09'))
# 2016 q4
q4_16 <- all_enc %>% filter(str_detect(DX_FROM_DATE, '2016-10|2016-11|2016-12'))
```

count number of excerbations in each quarter

```{r}
quarter = c("11-1", "11-2", "11-3", "11-4", "12-1", "12-2", "12-3", "12-4", "13-1", "13-2", "13-3", "13-4", "14-1", "14-2", "14-3", "14-4", "15-1", "15-2", "15-3", "15-4", "16-1", "16-2", "16-3", "16-4")
run_qrt = c (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 , 16, 17, 18, 19, 20, 21, 22, 23, 24)
counts = c(nrow(q1_11), nrow(q2_11), nrow(q3_11), nrow(q4_11), nrow(q1_12), nrow(q2_12), nrow(q3_12), nrow(q4_12), nrow(q1_13), nrow(q2_13), nrow(q3_13), nrow(q4_13),nrow(q1_14), nrow(q2_14), nrow(q3_14), nrow(q4_14), nrow(q1_15), nrow(q2_15), nrow(q3_15), nrow(q4_15),nrow(q1_16), nrow(q2_16), nrow(q3_16), nrow(q4_16))
asthma_count <- data.frame(quarter, run_qrt, counts)
```

plot total asthma exacerbations over time

```{r}
plot(asthma_count$run_qrt, asthma_count$counts)
```

The plot of all asthma encounters over time is not consistent with the severe encounters(hosp and ER visits). This makes me wonder if the reporting method for encounter type changed during the study period. As mentioned above, I need to talk with Jen to obtain a data dictionary that defines the encounter types if available.

export csv

```{r}
write.csv(asthma_count,"asthma_count_total.csv", row.names = TRUE)
write.csv(all_enc,"asthma_allenc.csv", row.names = TRUE)
```

Appendix A.2 PM 2.5 Data Cleaning

The following code was used to clean the PM 2.5 raster data.

load libraries

```{r}
##install.packages("sf")
##install.packages("rgeos")
##install.packages("proj4")
##install.packages("rgdal")
##install.packages("exactextractr")
library(exactextractr)
library(rgdal)
library(proj4)
```

```
library(rgeos)
library(sf)
library(maptools)
library(raster)
```

load data

raster maps

These are the PM2.5 maps for Jan, Apr, Jul, and Oct for 2011-2016.

2011

```{r}
jan2011 <- raster("D:/Maggie/Capstone/2011 pm2.5/V4NA03_PM25_NA_201101_201101-RH35-NoNegs.asc")
apr2011 <- raster("D:/Maggie/Capstone/2011 pm2.5/V4NA03_PM25_NA_201104_201104-RH35-NoNegs.asc")
jul2011 <- raster("D:/Maggie/Capstone/2011 pm2.5/V4NA03_PM25_NA_201107_201107-RH35-NoNegs.asc")
oct2011 <- raster("D:/Maggie/Capstone/2011 pm2.5/V4NA03_PM25_NA_201110_201110-RH35-NoNegs.asc")
```

2012

```{r}
jan2012 <- raster("D:/Maggie/Capstone/2012 pm2.5/V4NA03_PM25_NA_201201_201201-RH35-NoNegs.asc")
```

```
apr2012 <- raster("D:/Maggie/Capstone/2012 pm2.5/V4NA03_PM25_NA_201204_201204-RH35-NoNegs.asc")
jul2012 <- raster("D:/Maggie/Capstone/2012 pm2.5/V4NA03_PM25_NA_201207_201207-RH35-NoNegs.asc")
oct2012 <- raster("D:/Maggie/Capstone/2012 pm2.5/V4NA03_PM25_NA_201210_201210-RH35-NoNegs.asc")
```

2013

```{r}
jan2013 <- raster("D:/Maggie/Capstone/2013 pm2.5/V4NA03_PM25_NA_201301_201301-RH35-NoNegs.asc")
apr2013 <- raster("D:/Maggie/Capstone/2013 pm2.5/V4NA03_PM25_NA_201304_201304-RH35-NoNegs.asc")
jul2013 <- raster("D:/Maggie/Capstone/2013 pm2.5/V4NA03_PM25_NA_201307_201307-RH35-NoNegs.asc")
oct2013 <- raster("D:/Maggie/Capstone/2013 pm2.5/V4NA03_PM25_NA_201310_201310-RH35-NoNegs.asc")
```

2014

```{r}
jan2014 <- raster("D:/Maggie/Capstone/2014 pm2.5/V4NA03_PM25_NA_201401_201401-RH35-NoNegs.asc")
```

```
apr2014 <- raster("D:/Maggie/Capstone/2014 pm2.5/V4NA03_PM25_NA_201404_201404-RH35-NoNegs.asc")
jul2014 <- raster("D:/Maggie/Capstone/2014 pm2.5/V4NA03_PM25_NA_201407_201407-RH35-NoNegs.asc")
oct2014 <- raster("D:/Maggie/Capstone/2014 pm2.5/V4NA03_PM25_NA_201410_201410-RH35-NoNegs.asc")
```

2015

```{r}
jan2015 <- raster("D:/Maggie/Capstone/2015 pm2.5/V4NA03_PM25_NA_201501_201501-RH35-NoNegs.asc")
apr2015 <- raster("D:/Maggie/Capstone/2015 pm2.5/V4NA03_PM25_NA_201504_201504-RH35-NoNegs.asc")
jul2015 <- raster("D:/Maggie/Capstone/2015 pm2.5/V4NA03_PM25_NA_201507_201507-RH35-NoNegs.asc")
oct2015 <- raster("D:/Maggie/Capstone/2015 pm2.5/V4NA03_PM25_NA_201510_201510-RH35-NoNegs.asc")
```

2016

```{r}
jan2016 <- raster("D:/Maggie/Capstone/2016 pm2.5/V4NA03_PM25_NA_201601_201601-RH35-NoNegs.asc")
```

```
apr2016 <- raster("D:/Maggie/Capstone/2016 pm2.5/V4NA03_PM25_NA_201604_201604-RH35-NoNegs.asc")
jul2016 <- raster("D:/Maggie/Capstone/2016 pm2.5/V4NA03_PM25_NA_201607_201607-RH35-NoNegs.asc")
oct2016 <- raster("D:/Maggie/Capstone/2016 pm2.5/V4NA03_PM25_NA_201610_201610-RH35-NoNegs.asc")
```

beaver county tract shape file

```{r}
#read-in the polygon shapefile for Beaver county
beavertract <- st_read("D:/Maggie/Capstone/beaver.shp")
plot(beavertract)
```

plot pm2.5 with beaver county overlaid

```{r}
crs(jan2011) <- "+proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +towgs84=0,0,0"
# crop the pm2.5 raster using the vector extent
#jul2011_clip <- crop(jul2011, beavertract)
#plot(jul2011_clip, main = "Cropped Jul 2011")
# add shapefile on top of the existing raster
#plot(beavertract, col = NA, add = TRUE)
```

This map shows the average PM2.5 values for Jul 2011. THe dark green area in the center is where the zinc smelter was and high levels of PM2.5 are seen there.

Zonal statistics

Zonal statistics is a tool that is used to aggregate raster data by the parameters of some shape file. I will be averaging the PM2.5 value in each county tract in beaver county for each quarterly PM2.5 map.

2011

```{r}
# Calculate vector of mean 2011 Pm2.5 value for each tract
beavertract$mean_jan11 <- exact_extract(jan2011, beavertract, 'mean')
beavertract$mean_jul11 <- exact_extract(jul2011, beavertract, 'mean')
beavertract$mean_apr11 <- exact_extract(apr2011, beavertract, 'mean')
beavertract$mean_oct11 <- exact_extract(oct2011, beavertract, 'mean')
```

2012

```{r}
# Calculate vector of mean 2012 Pm2.5 value for each tract
beavertract$mean_jan12 <- exact_extract(jan2012, beavertract, 'mean')
beavertract$mean_jul12 <- exact_extract(jul2012, beavertract, 'mean')
beavertract$mean_apr12 <- exact_extract(apr2012, beavertract, 'mean')
beavertract$mean_oct12 <- exact_extract(oct2012, beavertract, 'mean')
```

2013

```{r}
# Calculate vector of mean 2013 Pm2.5 value for each tract
beavertract$mean_jan13 <- exact_extract(jan2013, beavertract, 'mean')
beavertract$mean_jul13 <- exact_extract(jul2013, beavertract, 'mean')
beavertract$mean_apr13 <- exact_extract(apr2013, beavertract, 'mean')
beavertract$mean_oct13 <- exact_extract(oct2013, beavertract, 'mean')
```

2014

```{r}
# Calculate vector of mean 2014 Pm2.5 value for each tract
beavertract$mean_jan14 <- exact_extract(jan2014, beavertract, 'mean')
beavertract$mean_jul14 <- exact_extract(jul2014, beavertract, 'mean')
beavertract$mean_apr14 <- exact_extract(apr2014, beavertract, 'mean')
beavertract$mean_oct14 <- exact_extract(oct2014, beavertract, 'mean')
```

2015

```{r}
# Calculate vector of mean 2015 Pm2.5 value for each tract
beavertract$mean_jan15 <- exact_extract(jan2015, beavertract, 'mean')
beavertract$mean_jul15 <- exact_extract(jul2015, beavertract, 'mean')
beavertract$mean_apr15 <- exact_extract(apr2015, beavertract, 'mean')
beavertract$mean_oct15 <- exact_extract(oct2015, beavertract, 'mean')
```

``

2016

```{r}
# Calculate vector of mean 2015 Pm2.5 value for each tract
beavertract$mean_jan16 <- exact_extract(jan2016, beavertract, 'mean')
beavertract$mean_jul16 <- exact_extract(jul2016, beavertract, 'mean')
beavertract$mean_apr16 <- exact_extract(apr2016, beavertract, 'mean')
beavertract$mean_oct16 <- exact_extract(oct2016, beavertract, 'mean')
```

Appendix A.3 Combining Data and Data Analysis

The following code was run on the cleaned data sets allenc_long for total asthma encounters and hosper_long for hospitalizations and ER visits for the statistical analysis.

Load library

```{r}
library(rgdal)
library(sp)
library(sf)
library(tidyverse)
#install.packages("devtools")
library(devtools)
```

```
#devtools::install_github("setzler/eventStudy/eventStudy")

library(eventStudy)

#install.packages("ivpack")

library(ivpack)

#install.packages("data.frame")

library(data.table)

#install.packages("zoo")

library(zoo)
```

#load dataset

```{r}
#beaver county shapefile

beaverpm25 <- st_read("beaverpm2_5.shp")

#beaver county csv

beaver25 <- read.csv("beaver25.csv")

#asthma exacerbations hospital and ER

hospercount <- read.csv("asthma_hosper_count.csv")

hosp_er<- read.csv("asthma_hosper_data.csv")

#all asthma encounters

allasthcount <- read.csv("asthma_count_total.csv")

all_enc <- read.csv("asthma_allenc.csv")
```

assign each encounter w county tract info

The xy points for all encounters and for hospital er data were mapped and assigned the data from the beaver tract shapefile based on location. The acrmap tool used was spatial join with one to one join

import new shapefiles

```{r}
#hospital and ER data
hosper_tracts <- read.csv("hosper_tract.csv")
#all encounters
allenc_tracts <- read.csv("allenc_tract.csv")
```

#make beaver pm 25 data long

```{r}
beaver25_2<- rename(beaver25,
mean_1 = mean_jan11, mean_2 = mean_apr11, mean_3 = mean_jul11, mean_4 = mean_oct11, mean_5 = mean_jan12, mean_6 = mean_apr12, mean_7 = mean_jul12, mean_8 = mean_oct12, mean_9 = mean_jan13, mean_10 = mean_apr13, mean_11 = mean_jul13, mean_12 = mean_oct13, mean_13 = mean_jan14, mean_14 = mean_apr14, mean_15 = mean_jul14, mean_16 = mean_oct14, mean_17 = mean_jan15, mean_18 = mean_apr15, mean_19 = mean_jul15, mean_20 = mean_oct15, mean_21 = mean_jan16, mean_22 = mean_apr16, mean_23 = mean_jul16, mean_24 = mean_oct16 )
beaver25_long <- gather(beaver25_2, quarter, pm25, mean_1:mean_24, factor_key=TRUE)
beaver25_long <- select(beaver25_long, quarter, TRACTCE20, pm25)
```

```
beaver25_long$quarter<- as.numeric(gsub("mean_", "", beaver25_long$quarter))

beaver25_long<- beaver25_long %>%

mutate(year = case_when(quarter == 1 ~ 2011,

                        quarter == 2 ~ 2011,

                        quarter == 3 ~ 2011,

                        quarter == 4 ~ 2011,

                        quarter == 5~ 2012,

                        quarter == 6~ 2012,

                        quarter == 7~ 2012,

                        quarter == 8~ 2012,

                        quarter == 9~ 2013,

                        quarter == 10~ 2013,

                        quarter == 11~ 2013,

                        quarter == 12~ 2013,

                        quarter == 13 ~ 2014,

                        quarter == 14 ~ 2014,

                        quarter == 15 ~ 2014,

                        quarter == 16 ~ 2014,

                        quarter == 17 ~ 2015,

                        quarter == 18 ~ 2015,

                        quarter == 19 ~ 2015,

                        quarter == 20 ~ 2015,

                        quarter == 21 ~ 2016,
```

```
                quarter == 22 ~ 2016,
                quarter == 23 ~ 2016,
                quarter == 24 ~ 2016,
                ))
```

sort by date exposures by date and assign proper pm value

hospital and ER data

```{r}
#select date, tract, and mean PM values
hosper_tract_clean <- select (hosper_tracts, DX_FROM_DA, TRACTCE20, mean_jan11:mean_oct16)
hosper_tract_clean<- rename(hosper_tract_clean,
 mean_1 = mean_jan11, mean_2 = mean_apr11, mean_3 = mean_jul11, mean_4 = mean_oct11,
 mean_5 = mean_jan12, mean_6 = mean_apr12, mean_7 = mean_jul12, mean_8 = mean_oct12,
 mean_9 = mean_jan13, mean_10 = mean_apr13, mean_11 = mean_jul13, mean_12 = mean_oct13,
mean_13 = mean_jan14, mean_14 = mean_apr14, mean_15 = mean_jul14, mean_16 = mean_oct14,
mean_17 = mean_jan15, mean_18 = mean_apr15, mean_19 = mean_jul15, mean_20 = mean_oct15,
mean_21 = mean_jan16, mean_22 = mean_apr16, mean_23 = mean_jul16, mean_24 = mean_oct16
    )
```

```

```{r}

# split into quarters

#2011

# 2011 q1

q1_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2011-01|2011-02|2011-03'))

# 2011 q2

q2_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2011-04|2011-05|2011-06'))

# 2011 q3

q3_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2011-07|2011-08|2011-09'))

# 2011 q4

q4_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2011-10|2011-11|2011-12'))

#2012

# 2012 q1

q5_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2012-01|2012-02|2012-03'))

# 2012 q2

q6_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2012-04|2012-05|2012-06'))

```
# 2012 q3
q7_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2012-07|2012-08|2012-09'))
# 2012 q4
q8_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2012-10|2012-11|2012-12'))
#2013
# 2013 q1
q9_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2013-01|2013-02|2013-03'))
# 2013 q2
q10_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2013-04|2013-05|2013-06'))
# 2013 q3
q11_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2013-07|2013-08|2013-09'))
# 2013 q4
q12_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2013-10|2013-11|2013-12'))
#2014
# 2014 q1
q13_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2014-01|2014-02|2014-03'))
```

```
# 2014 q2
q14_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2014-04|2014-05|2014-06'))
# 2014 q3
q15_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2014-07|2014-08|2014-09'))
# 2014 q4
q16_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2014-10|2014-11|2014-12'))
#2015
# 2015 q1
q17_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2015-01|2015-02|2015-03'))
# 2015 q2
q18_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2015-04|2015-05|2015-06'))
# 2015 q3
q19_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2015-07|2015-08|2015-09'))
# 2015 q4
q20_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2015-10|2015-11|2015-12'))
#2016
```

```
# 2016 q1
q21_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2016-01|2016-02|2016-03'))
# 2016 q2
q22_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2016-04|2016-05|2016-06'))
# 2016 q3
q23_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2016-07|2016-08|2016-09'))
# 2016 q4
q24_hosper <- hosper_tract_clean %>% filter(str_detect(DX_FROM_DA, '2016-10|2016-11|2016-12'))
```

load in census tract info data for rate

```{r}
pop11 <- read.csv("D:/Maggie/Capstone/2011census.csv.csv")
pop12 <- read.csv("D:/Maggie/Capstone/2012census.csv.csv")
pop13 <- read.csv("D:/Maggie/Capstone/2013census.csv.csv")
pop14 <- read.csv("D:/Maggie/Capstone/2014census.csv.csv")
pop15 <- read.csv("D:/Maggie/Capstone/2015census.csv.csv")
pop16 <- read.csv("D:/Maggie/Capstone/2016census.csv.csv")
```

```{r}
```

```
#add year to data
pop11$year <-  2011
pop12$year <- 2012
pop13$year <-  2013
pop14$year <- 2014
pop15$year <-  2015
pop16$year <- 2016
#merge all years
popallyears <- full_join(pop11, pop12) %>% full_join(pop13) %>% full_join(pop14) %>% full_join(pop15) %>% full_join(pop16)
#make popcount numeric
popallyears$B01003_001E <- as.numeric(popallyears$B01003_001E)
```

```{r}
#trim last 6 digits
n_last <- 6                         # Specify number of characters to extract
popallyears$TRACTCE20<-as.numeric(substr(popallyears$GEO_ID,
nchar(popallyears$GEO_ID) - n_last + 1, nchar(popallyears$GEO_ID)))
```

cleaning loop

```{r}
#loop for cleaning
clean <-  function(filename) {
```

```
# keep only pm for correct quarter
temp2 <- deparse(substitute(filename))
if (str_detect(temp2, "q1_")) {i = 1}
else if (str_detect(temp2, "q2_")) {i = 2}
else if (str_detect(temp2, "q3_")) {i = 3}
else if (str_detect(temp2, "q4_")) {i = 4}
else if (str_detect(temp2, "q5_")) {i = 5}
else if (str_detect(temp2, "q6_")) {i = 6}
else if (str_detect(temp2, "q7_")) {i = 7}
else if (str_detect(temp2, "q8_")) {i = 8}
else if (str_detect(temp2, "q9_")) {i = 9}
else if (str_detect(temp2, "q10_")) {i = 10}
else if (str_detect(temp2, "q11_")) {i = 11}
else if (str_detect(temp2, "q12_")) {i = 12}
else if (str_detect(temp2, "q13_")) {i = 13}
else if (str_detect(temp2, "q14_")) {i = 14}
else if (str_detect(temp2, "q15_")) {i = 15}
else if (str_detect(temp2, "q16_")) {i = 16}
else if (str_detect(temp2, "q17_")) {i = 17}
else if (str_detect(temp2, "q18_")) {i = 18}
else if (str_detect(temp2, "q19_")) {i = 19}
else if (str_detect(temp2, "q20_")) {i = 20}
else if (str_detect(temp2, "q21_")) {i = 21}
```

```
else if (str_detect(temp2, "q22_")) {i = 22}

else if (str_detect(temp2, "q23_")) {i = 23}

else if (str_detect(temp2, "q24_")) {i = 24}

filename <- select(filename, TRACTCE20, paste0( "mean_", i))

#get rid of data outside of beaver county

filename <- na.omit(filename)

#rename to PM

filename <- rename(filename, PM = paste0("mean_", i) )

#make column of quarter

filename$quarter <- i

# count up duplicates

filename <- filename %>%

  group_by(across(everything())) %>%

  mutate(n = n())

#remove duplicate rows

filename <- distinct (filename)

# add year

x<- i/4

filename$year <- filename %>% mutate(year = 2010 + ceiling(x))

}
```

```{r}

#clean all data with clean function
```

```
q1_hospclean<-clean(q1_hosper)

q2_hospclean<-clean(q2_hosper)

q3_hospclean<-clean(q3_hosper)

q4_hospclean<-clean(q4_hosper)

q5_hospclean<-clean(q5_hosper)

q6_hospclean<-clean(q6_hosper)

q7_hospclean<-clean(q7_hosper)

q8_hospclean<-clean(q8_hosper)

q9_hospclean<-clean(q9_hosper)

q10_hospclean<-clean(q10_hosper)

q11_hospclean<-clean(q11_hosper)

q12_hospclean<-clean(q12_hosper)

q13_hospclean<-clean(q13_hosper)

q14_hospclean<-clean(q14_hosper)

q15_hospclean<-clean(q15_hosper)

q16_hospclean<-clean(q16_hosper)

q17_hospclean<-clean(q17_hosper)

q18_hospclean<-clean(q18_hosper)

q19_hospclean<-clean(q19_hosper)

q20_hospclean<-clean(q20_hosper)

q21_hospclean<-clean(q21_hosper)

q22_hospclean<-clean(q22_hosper)

q23_hospclean<-clean(q23_hosper)
```

q24_hospclean<-clean(q24_hosper)

```

```{r}

#merge all data

hosper_merged <- full_join(q1_hospclean, q2_hospclean) %>% full_join(q3_hospclean) %>% full_join(q4_hospclean) %>% full_join(q5_hospclean) %>% full_join(q6_hospclean) %>% full_join(q7_hospclean) %>% full_join(q8_hospclean) %>% full_join(q9_hospclean) %>% full_join(q10_hospclean) %>% full_join(q11_hospclean) %>% full_join(q12_hospclean) %>% full_join(q13_hospclean) %>% full_join(q14_hospclean) %>% full_join(q15_hospclean) %>% full_join(q16_hospclean) %>% full_join(q17_hospclean) %>% full_join(q18_hospclean) %>% full_join(q19_hospclean) %>% full_join(q20_hospclean) %>% full_join(q21_hospclean) %>% full_join(q22_hospclean) %>% full_join(q23_hospclean) %>% full_join(q24_hospclean)

```

clean merged data

```{r}

#merge long pm data with clean hosper_merged

beaver25_long_pop<- full_join(beaver25_long, popallyears)

#merge population data

hosper_long <- full_join(hosper_merged, beaver25_long_pop)

```

```{r}

hosper_long <- select(hosper_long, TRACTCE20, quarter, n, year, pm25, B01003_001E)

```
```

```{r}

#make NAs 0

hosper_long[is.na(hosper_long)] <- 0

```

```{r}

hosper_long <- hosper_long %>% filter((quarter) != 0)

#temporary populTIONS OF 0

hosper_long <- hosper_long %>% filter((B01003_001E) != 0)

```

make n a rate

```{r}

hosper_long$rate <- (hosper_long$n/hosper_long$B01003_001E)*10000

```

rate of asthma hospitalization or ER visit per 10,000 people in each county tract during each quarter

#exposure time period

```{r}

hosper_long$post <- ifelse(hosper_long$quarter>=15, 1, 0)

```

ALL ENC

```{r}

#select date, tract, and mean PM values

allenc_tract_clean <- select (allenc_tracts, DX_FROM_DA, TRACTCE20, mean_jan11:mean_oct16)
```

```
allenc_tract_clean<- rename(allenc_tract_clean,
mean_1 = mean_jan11, mean_2 = mean_apr11, mean_3 = mean_jul11, mean_4 = mean_oct11,
mean_5 = mean_jan12, mean_6 = mean_apr12, mean_7 = mean_jul12, mean_8 = mean_oct12,
 mean_9 = mean_jan13, mean_10 = mean_apr13, mean_11 = mean_jul13, mean_12 = mean_oct13,
mean_13 = mean_jan14, mean_14 = mean_apr14, mean_15 = mean_jul14, mean_16 = mean_oct14,
mean_17 = mean_jan15, mean_18 = mean_apr15, mean_19 = mean_jul15, mean_20 = mean_oct15,
mean_21 = mean_jan16, mean_22 = mean_apr16, mean_23 = mean_jul16, mean_24 = mean_oct16)
```

```{r}
# split into quarters
#2011
# 2011 q1
q1_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2011-01|2011-02|2011-03'))
# 2011 q2
q2_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2011-04|2011-05|2011-06'))
# 2011 q3
```

```
q3_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2011-07|2011-08|2011-09'))
# 2011 q4
q4_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2011-10|2011-11|2011-12'))
#2012
# 2012 q1
q5_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2012-01|2012-02|2012-03'))
# 2012 q2
q6_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2012-04|2012-05|2012-06'))
# 2012 q3
q7_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2012-07|2012-08|2012-09'))
# 2012 q4
q8_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2012-10|2012-11|2012-12'))
#2013
# 2013 q1
q9_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2013-01|2013-02|2013-03'))
# 2013 q2
```

```
q10_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2013-04|2013-05|2013-
06'))
# 2013 q3
q11_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2013-07|2013-08|2013-
09'))
# 2013 q4
q12_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2013-10|2013-11|2013-
12'))
#2014
# 2014 q1
q13_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2014-01|2014-02|2014-
03'))
# 2014 q2
q14_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2014-04|2014-05|2014-
06'))
# 2014 q3
q15_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2014-07|2014-08|2014-
09'))
# 2014 q4
q16_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2014-10|2014-11|2014-
12'))
#2015
# 2015 q1
```

```
q17_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2015-01|2015-02|2015-
03'))
# 2015 q2
q18_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2015-04|2015-05|2015-
06'))
# 2015 q3
q19_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2015-07|2015-08|2015-
09'))
# 2015 q4
q20_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2015-10|2015-11|2015-
12'))
#2016
# 2016 q1
q21_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2016-01|2016-02|2016-
03'))
# 2016 q2
q22_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2016-04|2016-05|2016-
06'))
# 2016 q3
q23_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2016-07|2016-08|2016-
09'))
# 2016 q4
```

q24_allenc <- allenc_tract_clean %>% filter(str_detect(DX_FROM_DA, '2016-10|2016-11|2016-12'))

```

```{r}

#clean all data with clean function

q1_allencclean<-clean(q1_allenc)

q2_allencclean<-clean(q2_allenc)

q3_allencclean<-clean(q3_allenc)

q4_allencclean<-clean(q4_allenc)

q5_allencclean<-clean(q5_allenc)

q6_allencclean<-clean(q6_allenc)

q7_allencclean<-clean(q7_allenc)

q8_allencclean<-clean(q8_allenc)

q9_allencclean<-clean(q9_allenc)

q10_allencclean<-clean(q10_allenc)

q11_allencclean<-clean(q11_allenc)

q12_allencclean<-clean(q12_allenc)

q13_allencclean<-clean(q13_allenc)

q14_allencclean<-clean(q14_allenc)

q15_allencclean<-clean(q15_allenc)

q16_allencclean<-clean(q16_allenc)

q17_allencclean<-clean(q17_allenc)

q18_allencclean<-clean(q18_allenc)

```
q19_allencclean<-clean(q19_allenc)

q20_allencclean<-clean(q20_allenc)

q21_allencclean<-clean(q21_allenc)

q22_allencclean<-clean(q22_allenc)

q23_allencclean<-clean(q23_allenc)

q24_allencclean<-clean(q24_allenc)
```

```{r}
#merge all data

allenc_merged <- full_join(q1_allencclean, q2_allencclean) %>% full_join(q3_allencclean) %>% full_join(q4_allencclean) %>% full_join(q5_allencclean) %>% full_join(q6_allencclean) %>% full_join(q7_allencclean) %>% full_join(q8_allencclean) %>% full_join(q9_allencclean) %>% full_join(q10_allencclean) %>% full_join(q11_allencclean) %>% full_join(q12_allencclean) %>% full_join(q13_allencclean) %>% full_join(q14_allencclean) %>% full_join(q15_allencclean) %>% full_join(q16_allencclean) %>% full_join(q17_allencclean) %>% full_join(q18_allencclean) %>% full_join(q19_allencclean) %>% full_join(q20_allencclean) %>% full_join(q21_allencclean) %>% full_join(q22_allencclean) %>% full_join(q23_allencclean) %>% full_join(q24_allencclean)
```

clean merged data

```{r}
#merge long pm data with clean allenc_merged

beaver25_long_pop<- full_join(beaver25_long, popallyears)
```

#merge population data

allenc_long <- full_join(allenc_merged, beaver25_long_pop)

```

```{r}

allenc_long <- select(allenc_long, TRACTCE20, quarter, n, year, pm25, B01003_001E)

```

```{r}

#make NAs 0

allenc_long[is.na(allenc_long)] <- 0

```

```{r}

allenc_long <- allenc_long %>% filter((quarter) != 0)

#temporary populTIONS OF 0

allenc_long <- allenc_long %>% filter((B01003_001E) != 0)

```

make n a rate

```{r}

allenc_long$rate <- (allenc_long$n/allenc_long$B01003_001E)*10000

```

rate of all encounters per 10,000 people in each county tract during each quarter

#exposure time period

```{r}

allenc_long$post <- ifelse(allenc_long$quarter>=15, 1, 0)

```

Zinc smelter closed at the end May 2014

Statistical Analysis

make exposure zone

```{r}

allenc_long$exposed_final<- ifelse(allenc_long$TRACTCE20 == 603300 | allenc_long$TRACTCE20 == 603400 | allenc_long$TRACTCE20 == 603500 | allenc_long$TRACTCE20 == 603600 | allenc_long$TRACTCE20 == 603700 | allenc_long$TRACTCE20 == 604000 | allenc_long$TRACTCE20 == 604100 | allenc_long$TRACTCE20 == 604200 | allenc_long$TRACTCE20 == 604500 | allenc_long$TRACTCE20 == 604600 | allenc_long$TRACTCE20 == 604700 | allenc_long$TRACTCE20 == 601200 | allenc_long$TRACTCE20 == 601000 | allenc_long$TRACTCE20 == 603202 | allenc_long$TRACTCE20 == 604901 | allenc_long$TRACTCE20 == 601300 | allenc_long$TRACTCE20 == 601400 | allenc_long$TRACTCE20 == 601600 | allenc_long$TRACTCE20 == 602100 | allenc_long$TRACTCE20 == 602300 | allenc_long$TRACTCE20 == 602400 | allenc_long$TRACTCE20 == 602500 | allenc_long$TRACTCE20 == 605700 | allenc_long$TRACTCE20 == 605600 | allenc_long$TRACTCE20 == 603900 | allenc_long$TRACTCE20 == 605400 | allenc_long$TRACTCE20 == 605500 | allenc_long$TRACTCE20 == 604800 | allenc_long$TRACTCE20 == 601100 | allenc_long$TRACTCE20 == 605200 | allenc_long$TRACTCE20 == 602701 , 1, 0)

```

make exposure zone
```

```{r}
hosper_long$exposed_final<- ifelse(hosper_long$TRACTCE20 == 603300 | hosper_long$TRACTCE20 == 603400 | hosper_long$TRACTCE20 == 603500 | hosper_long$TRACTCE20 == 603600 | hosper_long$TRACTCE20 == 603700 | hosper_long$TRACTCE20 == 604000 | hosper_long$TRACTCE20 == 604100 | hosper_long$TRACTCE20 == 604200 | hosper_long$TRACTCE20 == 604500 | hosper_long$TRACTCE20 == 604600 | hosper_long$TRACTCE20 == 604700 | hosper_long$TRACTCE20 == 601200 | hosper_long$TRACTCE20 == 601000 | hosper_long$TRACTCE20 == 603202 | hosper_long$TRACTCE20 == 604901 | hosper_long$TRACTCE20 == 601300 | hosper_long$TRACTCE20 == 601400 | hosper_long$TRACTCE20 == 601600 | hosper_long$TRACTCE20 == 602100 | hosper_long$TRACTCE20 == 602300 | hosper_long$TRACTCE20 == 602400 | hosper_long$TRACTCE20 == 602500 | hosper_long$TRACTCE20 == 605700 | hosper_long$TRACTCE20 == 605600 | hosper_long$TRACTCE20 == 603900 | hosper_long$TRACTCE20 == 605400 | hosper_long$TRACTCE20 == 605500 | hosper_long$TRACTCE20 == 604800 | hosper_long$TRACTCE20 == 601100 | hosper_long$TRACTCE20 == 605200 | hosper_long$TRACTCE20 == 602701 , 1, 0)
```

MAKE SEASONAL EFFECT

```{r}
allenc_long$summer <- ifelse (allenc_long$quarter == 3 | allenc_long$quarter == 7 | allenc_long$quarter == 11 |allenc_long$quarter == 15 | allenc_long$quarter == 19 | allenc_long$quarter == 23, 1, 0 )
```

```

```{r}
hosper_long$summer <- ifelse (hosper_long$quarter == 3 | hosper_long$quarter == 7 | hosper_long$quarter == 11 |hosper_long$quarter == 15 | hosper_long$quarter == 19 | hosper_long$quarter == 23, 1, 0 )
```

1st stage preliminary model

```{r}
hosper_long$exposed_final <- as.character(hosper_long$exposed_final)
# relationship of PM2.5 and time
 ggplot(hosper_long, aes(quarter, pm25, color = exposed_final)) +
 stat_summary(geom = "errorbar") + stat_summary( geom = "point") + xlab("Quarter from 2011-2016") + ylab("PM 2.5 (ug/m3)") + geom_vline(xintercept = 15)

# relationship of PM2.5 and time
ggplot(hosper_long, aes(quarter, pm25, color = exposed_final)) +
geom_vline(xintercept = 15) + geom_smooth(aes(color = exposed_final), method = lm, formula = y ~ splines::bs(x, df = 2, degree = 1, knots = 15)) + xlab("Quarter from 2011- 2016") + ylab("PM 2.5 (ug/m3)")
```

HOSPITAL AND ER VISITS

```{r}

```
hosper_long$exposed_final <- as.character(hosper_long$exposed_final)
relationship of hosp er and time
ggplot(hosper_long, aes(quarter, rate, color = exposed_final)) +
stat_summary(geom = "errorbar") + stat_summary(geom = "point") + xlab("Quarter from 2011-2016") + ylab("Asthma Hospitalization or ER vist (per 10,000 people)") + geom_vline(xintercept = 15)
```

```{r}
hosper_long$exposed_final <- as.character(hosper_long$exposed_final)
# relationship of asthma rate and time trends
ggplot(hosper_long, aes(quarter, rate, color = exposed_final)) +
geom_vline(xintercept = 15) + geom_smooth(aes(color = exposed_final), method = lm,
formula = y ~ splines::bs(x, df = 2, degree = 1, knots = 15)) +  xlab("Quarter from 2011- 2016") + ylab("Asthma Hospitalization or ER vist (per 10,000 people)")
```

ITT

```{r}
#hosper
fit_did_itt1_final <- lm(rate ~ exposed_final + post + exposed_final*post +summer, data = hosper_long)
summary (fit_did_itt1_final)
```

parallel trends test

```{r}

#hosper parallel

fit_did_parallel1_final <- lm(rate ~ factor(quarter)*exposed_final + summer , data = hosper_long)

summary (fit_did_parallel1_final)

```

1SLS

```{r}

SLS1.1_final <- lm(pm25 ~ exposed_final + post + exposed_final*post + summer, data = hosper_long)

summary(SLS1.1_final)

pred_pm1_final = predict(SLS1.1_final, type = "response")

```

2SLS

```{r}

SLS2.1_final <- lm(rate ~ pred_pm1_final + post + exposed_final + summer, data = hosper_long)

summary(SLS2.1_final)

```

instrument test

```{r}

#IV model with covariates

ivmodel1_final = ivreg(formula = rate ~ post + exposed_final + pm25, instruments = ~ post + exposed_final + post*exposed_final, data = hosper_long)
```

```
# Notice that we use the robust sandwich covariance estimator here
summary(ivmodel1_final, vcov = sandwich, diagnostics = TRUE)
```

ALL ENCOUNTERS

```{r}
allenc_long$exposed_final <- as.character(allenc_long$exposed_final)
# relationship of all encounters and time
ggplot(allenc_long, aes(quarter, rate, color = exposed_final)) +

stat_summary(geom = "errorbar") + stat_summary( geom = "point") + xlab("Quarter from 2011-2016") + ylab("Total Asthma Encounter Rate (per 10,000 people)") + geom_vline(xintercept = 15)
```

```{r}
# relationship of asthma rate and time trends
ggplot(allenc_long, aes(quarter, rate, color = exposed_final)) +
geom_vline(xintercept = 15) + geom_smooth(aes(color = exposed_final), method = lm, formula = y ~ splines::bs(x, df = 2, degree = 1, knots = 15 )) + xlab("Quarter from 2011- 2016") + ylab("Total Asthma Encounter Rate (per 10,000 people)")
```

ITT

```{r}
```

#all encounters

fit_did_itt2_final <- lm(rate ~ exposed_final + post + exposed_final*post +summer , data = allenc_long)

summary (fit_did_itt2_final)

```

# test parallel trends

```{r}

#all encounters

fit_did_parallel2_final <- lm(rate ~ factor(quarter)*exposed_final +summer , data = allenc_long)

summary (fit_did_parallel2_final)

```

1SLS

```{r}

SLS1.2_final <- lm(pm25 ~ exposed_final + post + exposed_final*post + summer, data = allenc_long)

summary(SLS1.2_final)

pred_pm2_final = predict(SLS1.2_final, type = "response")

```

#2SLS

```{r}

SLS2.2_final <- lm(rate ~ pred_pm2_final + post + exposed_final + summer, data = allenc_long)

summary(SLS2.2_final)

```
```

instrument test

```{r}

#IV model with covariates

ivmodel2_final = ivreg(formula = rate ~ post + exposed_final + pm25, instruments = ~ post + exposed_final + post*exposed_final, data = allenc_long)

# Notice that we use the robust sandwich covariance estimator here

summary(ivmodel2_final, vcov = sandwich, diagnostics = TRUE)

```

Appendix B Parallel Pre-trends Test

The model below was used to assess if the pre-trends in the intention to treat model were parallel. The δ coefficient values represent the interaction of each quarter with the exposed variable before the closure of the zinc smelter. Statistically insignificant coefficient values mean that there is no difference in the slopes for the exposed and unexposed groups at each quarter and the trends in the pre-closure period are parallel.

$$asthma\ outcomes_{zt} = \beta_0 + \sum_{t<15} \theta_t(Quarter_t) + \sum_{t\geq 15} \gamma_t(Quarter_t = t) + \boldsymbol{\sum_{t<15} \delta_t(Quarter_t = t) * Exposed_z} + \sum_{t\geq 15} \eta_t(Age_t = t) * Exposed_z + \beta_1 * summer_t + \varepsilon_{zt}$$ **Equation 5**

Table 9 Appendix: Parallel Pre-trends Test for Total Athsma Encounters

Coefficients, δ	Estimate	Std. Error	t value	Pr(>|t|)
factor(quarter)2:exposed_final1	1.75367	3.03719	0.577	0.5638
factor(quarter)3:exposed_final1	0.54409	3.03719	0.179	0.8579
factor(quarter)4:exposed_final1	2.24664	3.03719	0.740	0.4596
factor(quarter)5:exposed_final1	2.06250	3.03719	0.679	0.4972
factor(quarter)6:exposed_final1	2.50843	3.03719	0.826	0.4090
factor(quarter)7:exposed_final1	-1.14967	3.03719	-0.379	0.7051
factor(quarter)8:exposed_final1	1.48559	3.03719	0.489	0.6248
factor(quarter)9:exposed_final1	-2.25356	3.03719	-0.742	0.4582
factor(quarter)10:exposed_final1	3.47729	3.03719	1.145	0.2525
factor(quarter)11:exposed_final1	1.79221	3.03719	0.590	0.5552

factor(quarter)12:exposed_final1	-0.38539	3.03719	-0.127	0.8991
factor(quarter)13:exposed_final1	-0.27191	3.03719	-0.090	0.9287
factor(quarter)14:exposed_final1	0.14577	3.03719	0.048	0.9617

Table 10 Appendix: Parallel Pre-trends Test for Hospitalizations and ER Visits

Coefficients, δ	Estimate	Std. Error	t value	Pr(>|t|)
factor(quarter)2:exposed_final1	0.260744	0.724309	0.360	0.7189
factor(quarter)3:exposed_final1	0.124802	0.724309	0.172	0.8632
factor(quarter)4:exposed_final1	0.416923	0.724309	0.576	0.5650
factor(quarter)5:exposed_final1	0.366679	0.724309	0.506	0.6128
factor(quarter)6:exposed_final1	0.194035	0.724309	0.268	0.7888
factor(quarter)7:exposed_final1	-0.430691	0.724309	-0.595	0.5522
factor(quarter)8:exposed_final1	-0.012451	0.724309	-0.017	0.9863
factor(quarter)9:exposed_final1	-0.759963	0.724309	-1.049	0.2943
factor(quarter)10:exposed_final1	0.628089	0.724309	0.867	0.3860
factor(quarter)11:exposed_final1	0.865873	0.724309	1.195	0.2322
factor(quarter)12:exposed_final1	-0.113534	0.724309	-0.157	0.8755
factor(quarter)13:exposed_final1	0.286066	0.724309	0.395	0.6930
factor(quarter)14:exposed_final1	-0.426984	0.724309	-0.590	0.5556
factor(quarter)15:exposed_final1	-0.141932	0.724309	-0.196	0.8447

The tables above show that the δ coefficient values are not statistically significant (p-values highlighted in yellow) and the trends are parallel for both asthma outcomes.

Appendix C Time Varying Exposure

This appendix shows the results using a time varying exposure model described in the discussion.

Appendix C.1 Exploratory Analysis

Exploratory analysis was done to assess temporal trends of both PM 2.5 and asthma exacerbations in both exposure groups.

Appendix C.1.1 First stage Relationship

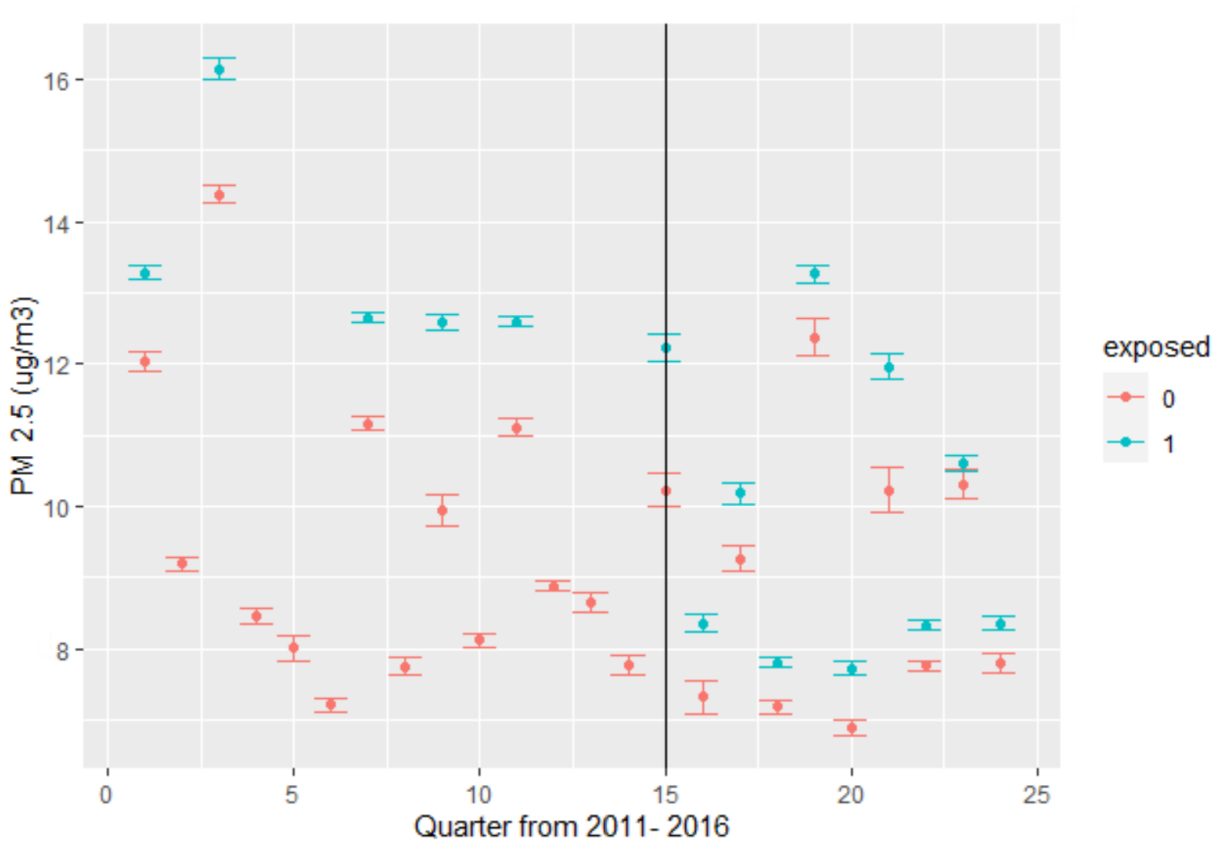

Figure 10 Appendix: Average PM 2.5 for exposed and unexposed tracts from 2011 - 2016

The graph above displays the changes in average PM 2.5 for county tracts in each exposure group over time from 2011 – 2016. An exposure value of 0 corresponds to unexposed county tracts and an exposure value of 1 corresponds to exposed county tracts.

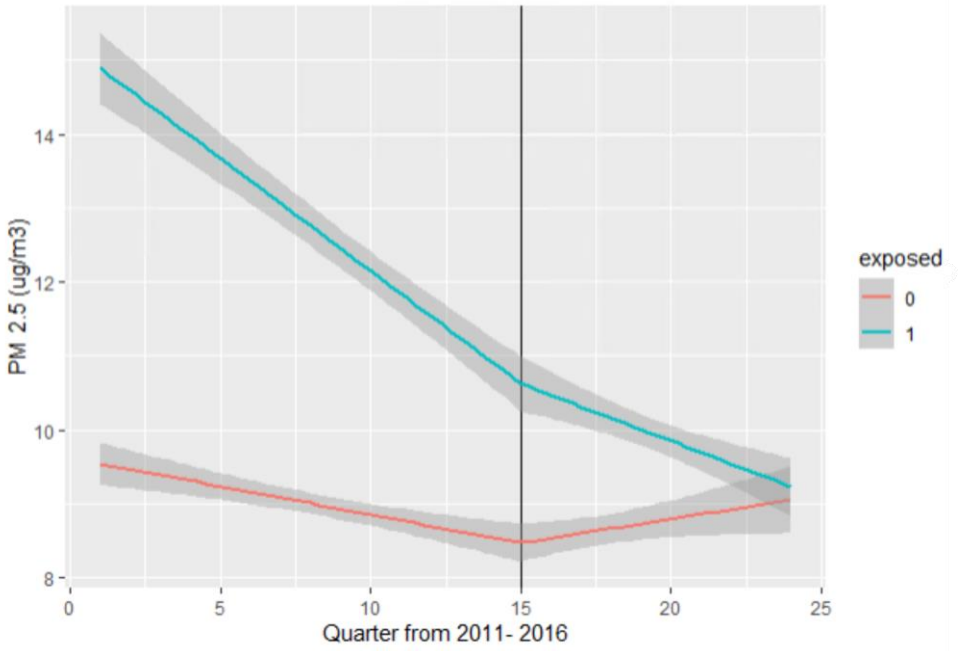

Figure 11 Appendix: Temporal Trends of PM 2.5 for exposued and enexposed groups

The graph above displays the temporal trend in PM 2.5 for county tracts in each exposure groups over time from 2011 – 2016. An exposure value of 0 corresponds to unexposed county tracts and an exposure value of 1 corresponds to exposed county tracts.

Figures 10 and 11 visually represent the first stage of the Two Stage Least Squares model. There is an overall decrease in PM 2.5 over the study period. In the post-closure period, PM 2.5 is decreasing in the exposure region and increasing slightly in the unexposed region.

Appendix C.1.2 Exploratory Analysis

Exploratory analysis was done for each outcome.

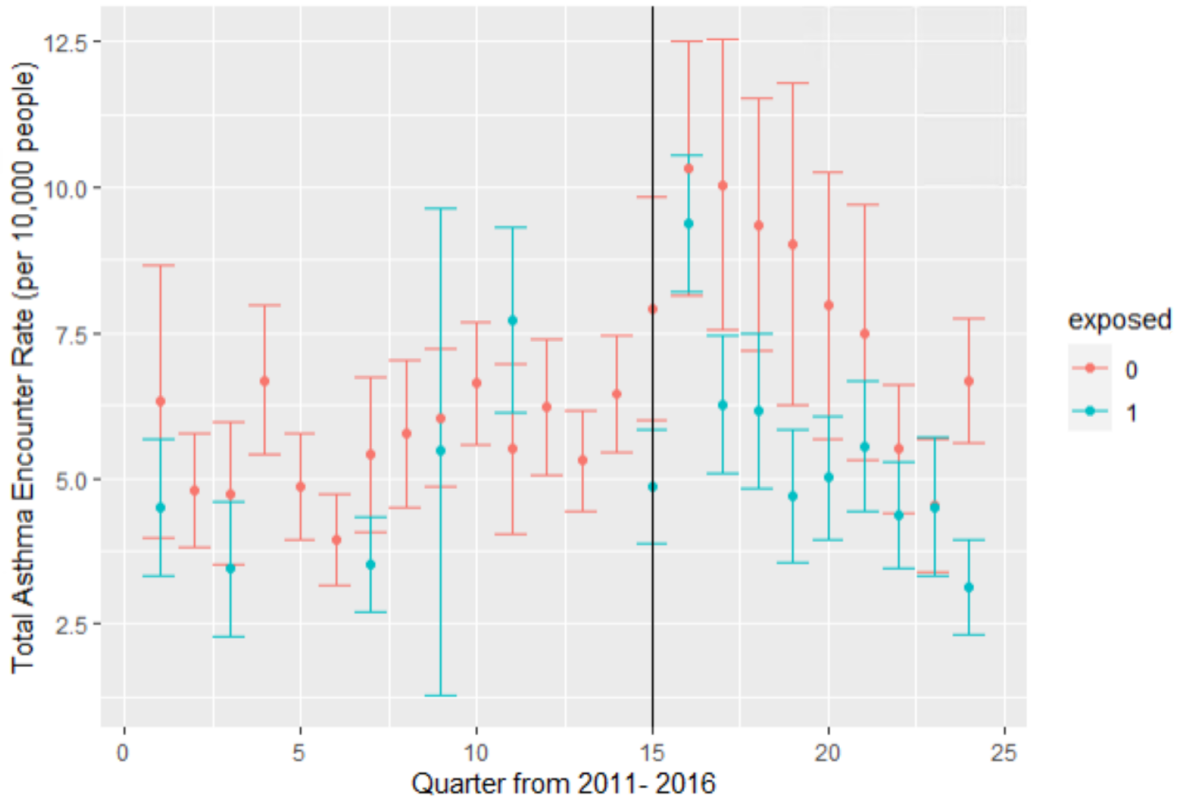

Figure 12 Appendix: Total Asthma Encounters from 2011 - 2016 in exposed and unexposed tracts

The graph above visualizes the relationship between rates of total asthma encounters and time by quarters for the two exposure groups. Rates of total asthma encounters appear to be higher in the unexposed group compared to the exposed group.

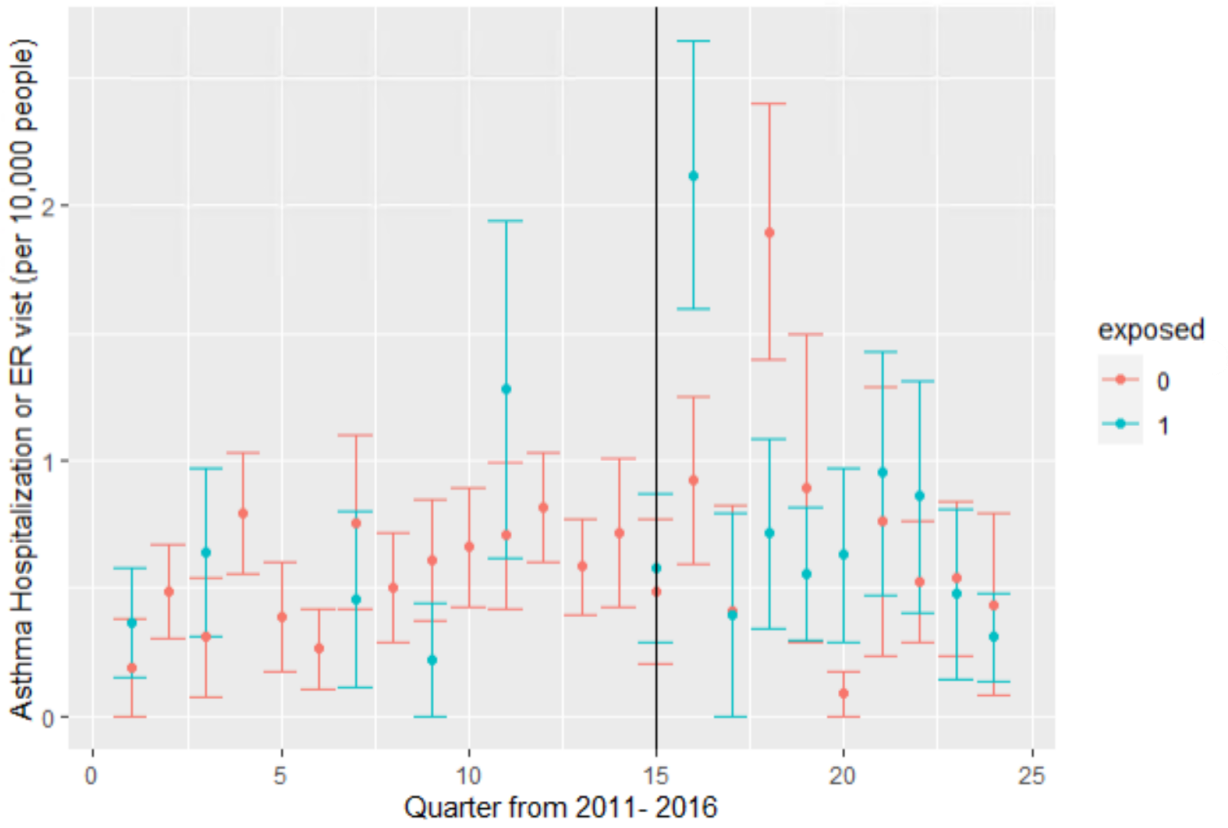

Figure 13 Appendix: Hospitalizations and ER visits from 2011 - 2016 in exposed and unexposed tracts

The graph above visualizes an event study model of the relationship between rates of hospital and emergency room visits and time by quarters for the two exposure groups.

The difference in difference analysis tells us if the differential change between the two groups is significant for each outcome. In the pre-treatment period some quarters do not have a measurement for the exposed group. This is because no county tracts in the at-risk zone had a PM 2.5 value greater than 12 $\mu g/m^3$during that quarter.

Appendix C.2 Difference in Difference Analysis

A difference in difference analysis was performed to assess the differential change in asthma rates before and after the zinc smelter closure.

Appendix C.2.1 Intention to treat model

An intention to treat model was performed to assess the effect of the zinc smelter closure on asthma encounters.

Table 11 Appendix: Estimates for Intention to treat model for Total Asthama Encounters

Coefficient	Estimate	P-value
Intercept, β_0	5.6336	< 2e-16 ***
Exposed, β_1	-1.0184	0.179004
Post, β_2	2.2568	0.000311 ***
Exposure Post interaction, β_3	-1.4815	0.147039

*** indicates statistical significance at the 0.05 level

The difference in difference estimator is represented by β_3. For rates of total asthma encounters, the difference in difference estimator is -1.48 and has a p-value of 0.147.

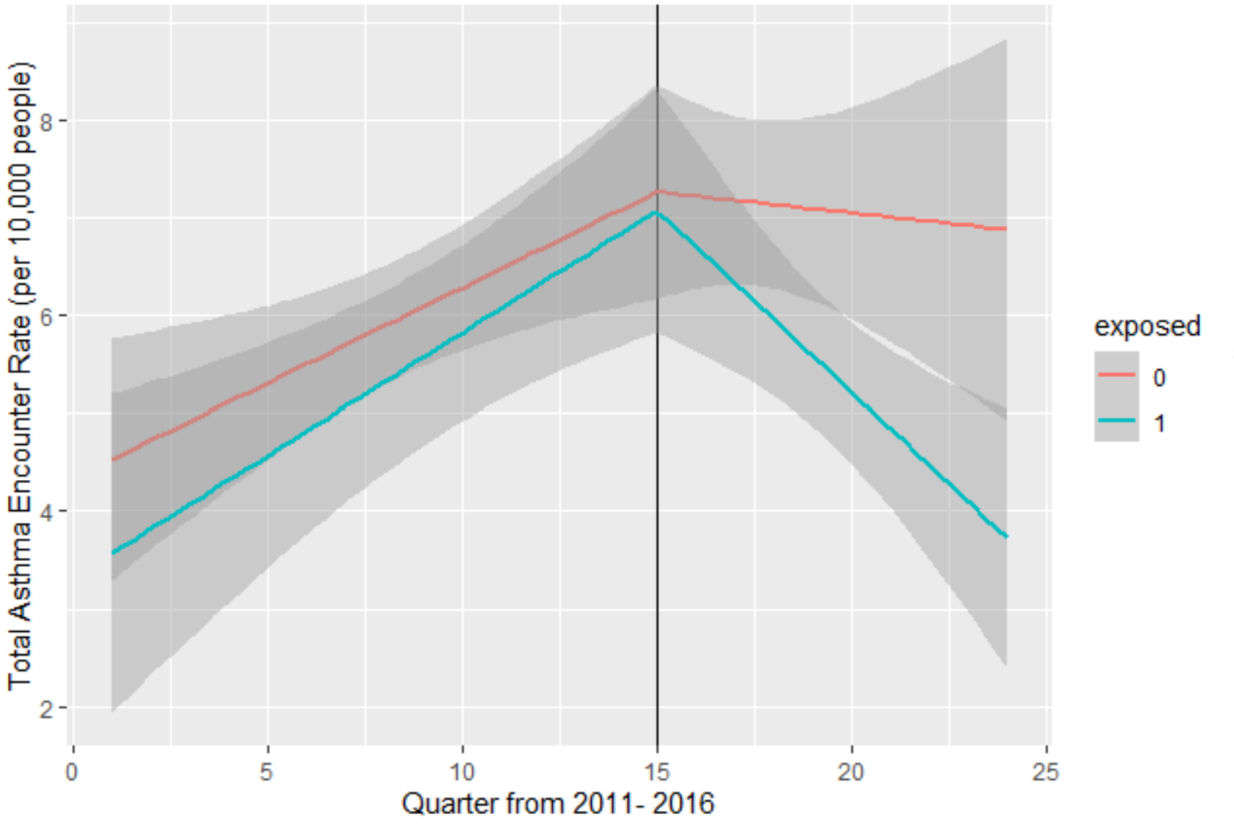

Figure 14 Appendix: Difference in Differnce for Total Asthma Encounters

The graph above displays the trends in average rate of total asthma encounters in the exposed and unexposed tracts over time. The rates of total asthma encounters are increasing in the pre-closure period in a similar trajectory for both groups. After the closure, total asthma rates are decreasing in both exposure groups. This is consistent with the results from the event study graph above. The rate of decrease is higher for the exposed group compared to the unexposed group. This is consistent with the difference in difference estimator, which is negative and therefore represents a greater decrease in the exposed groups compared to the unexposed group after the closure.

Table 12 Appendix: Estimates for Intention to Treat Model for Hospitalizations and Emergency Room Visits

Coefficient	Estimate	P-value
Intercept, β_0	0.575232	3.98e-15 ***
Exposed, β_1	0.009414	0.958
Post, β_2	0.121969	0.410
Exposure Post interaction, β_3	0.052861	0.827

*** indicates statistical significance at the 0.05 level

For rates hospitalizations and ER visits, the difference in difference estimator is 0.0529 and has a p-value of 0.827.

Figure 15 Appendix: Difference in Differnce for Hospitalizations and Emergency Room visits

The graph above displays the trends in average rate of hospitalizations and emergency room visits in the exposed and unexposed tracts over time. The rates of hospitalizations and emergency room visits are increasing in the pre-closure period and decreasing in the post exposure period. However, the pre-closure trends do not appear parallel so the difference in difference model must be adjusted. This may explain the positive difference in difference value of 0.0529 and low statistical significance.

Appendix C.3 Instrumental Variable Analysis

An instrumental variable analysis was performed to assess the causal effects of the zinc smelter closure on asthma rates through a change in air pollution.

Appendix C.3.1 First Stage Model

The first stage model represents the first part of the causal pathway – the effect of the zinc smelter closure on PM 2.5 air pollution.

Table 13 Appendix: First Stage Model Estimates

Coefficient	Estimate	P-value
Intercept, β_0	8.95720	< 2e-16 ***
Exposed, β_1	4.82086	<2e-16 ***
Post, β_2	-0.02127	0.894
Exposure Post interaction, β_3	-3.87531	<2e-16 ***

*** indicates statistical significance at the 0.05 level

The differential change in PM 2.5 between the two exposure groups is -3.88. This is consistent with trends in Figure 2. The high significance of this value shows that the zinc smelter closure is a strong instrument.

Appendix C.3.2 Second Stage Model

The second stage model represents the second part of the causal pathway – the effect of air pollution on asthma rates. The predicted values of PM 2.5 from the first stage model were used in the model.

Table 14 Appendix: Second Stage Estiamtes for Total Asthma Encounters

Coefficient	Estimate	P-value
Intercept, β_0	2.2094	0.370838
Predicted PM 2.5, β_1	0.3823	0.147039
Post, β_2	2.2649	0.000319 ***
Exposure, β_3	-2.8614	0.001056 **

*** indicates statistical significance at the 0.05 level

The β_1 value represents the Local Average Treatment Effect (LATE). The LATE for total asthma encounters is 0.382 with a p-value of 0.147.

Table 15 Appendix: Second Stage Estimates for Hospitalizations and Emergency Room Visits

Coefficient	Estimate	P-value
Intercept, β_0	0.69741	0.234

Predicted PM 2.5, β_1	-0.01364	0.827
Post, β_2	0.12168	0.414
Exposure, β_3	0.07517	0.716

*** indicates statistical significance at the 0.05 level

The LATE for total hospitalizations and emergency room visits is 0.0136 with a p-value of 0.827.

Bibliography

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics.* Princeton: Princeton University Press.

Atmospheric Composition Analysis Group. (2022). *Surface PM2.5*. Retrieved from Washington University of St. Louis: https://sites.wustl.edu/acag/datasets/surface-pm2-5/

Casey, J. A., Su, J. G., Henneman, L. R., Zigler, C., Neophytou, A. M., Catalano, R., . . . Barrett, M. A. (2020). Improved asthma outcomes observed in the vicinity of coal power plant retirement, retrofit and conversion to natural gas. *Nature Energy*, 398-408.

Edginton, S., O'Sullivan, D. E., King, W. D., & Lougheed, M. D. (2021). The effect of acute outdoor air pollution on peak expiratory flow in individuals with asthma: A systematic review and meta-analysis. *Environmental Research*, 1102956.

Erqou, S., Clougherty, J. E., Olafiranye, O., Magnani, J. W., Aiyer, A., Tripathy, S., . . . Reise, S. E. (2018). Particulate Matter Air Pollution and Racial Differences . *Arteriosclerosis, Thrombosis, and Vascular Biology*, 935-942.

Gough, P. J. (2016, February 12). *Timeline: Horsehead Holding Corp.* Retrieved from Pittsburgh Business Times: https://www.bizjournals.com/pittsburgh/news/2016/02/12/time-line-horsehead-holding-corp.html

Son, J.-Y., Fong, K. C., Heo, S., Kim, H., Lim, C. C., & Bell, M. L. (2020). Reductions in mortality resulting from reduced air pollution levels due to. *Science of the Total Environment*, 141012.

United States Census Bureau. (2022). *Explore Census Data*. Retrieved from https://data.census.gov/cedsci/

United States Environmental Protection Agency. (2021, May 26). *Particulate Matter (PM) Basics*. Retrieved from EPA: https://www.epa.gov/pm-pollution/particulate-matter-pm-basics

United States Environmental Protection Agency. (2022, April 5). *NAAQS Table*. Retrieved from EPA: https://www.epa.gov/criteria-air-pollutants/naaqs-table

World Health Organization. (2021, September 22). *Ambient (outdoor) air pollution*. Retrieved from WHO: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

WTAE. (2013, October 31). *Beaver County zinc plant closing, 'cracker' plans uncertain*. Retrieved from Pittsburgh's Action News: https://www.wtae.com/article/beaver-county-zinc-plant-closing-cracker-plans-uncertain/7463571